

QUALITY CONTROL FOR DNA DATA STORAGE

CROSS-REFERENCE TO RELATED APPLICATION

[001] This application claims priority to U.S. Provisional Patent Application No. 63/481,747, filed on January 26, 2023, which is hereby incorporated by reference in its entirety. All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BACKGROUND

[002] DNA is a compelling data storage medium given its superior density, stability, energy-efficiency, and longevity compared to currently used electronic media. However, errors and ambiguities can be introduced or otherwise occur at or during various stages of storage, sequencing and sequencing-related operations and processes. Therefore, there is a need to develop methods to efficiently perform quality control of DNA.

SUMMARY

[003] A goal of DNA data storage can be to provide a long lasting backup, especially, for example, where other backups fail. In such cases, verifying proper synthesis and/or storage of a polynucleotide pools can be critical. However, these pools can be extremely large, and a typical full sequencing run can be very expensive and may not be scalable. In such cases, it can also be necessary to develop methods to quality control polynucleotide pools over time. Quality control of polynucleotide pools can comprise verifying the data integrity and/or quantifying DNA degradation. In some instances, quality control further comprises correcting a polynucleotide pool or a subset thereof, as needed. In this instances, the original data in the polynucleotide pool is not available unless the polynucleotide pool is fully decoded and/or sequenced.

[004] Provided herein are designs and implementations for quality control of a plurality of polynucleotides. The plurality of polynucleotides can store digital information (e.g., binary data). The quality control of the plurality of polynucleotides can be performed before, during, and/or after synthesis or storage of the plurality of polynucleotides. The quality control of the plurality of polynucleotides can be performed on any suitable synthesis or storage device, such as those described herein. The quality control can further comprise quality control polynucleotides and/or one or more codecs, as further described herein. The provided designs and implementations can provide cost-effective and/or scalable quality control of a plurality of polynucleotides.

[005] In one aspect, provided herein is a method for quality control (QC) of data polynucleotides, comprising: (i) providing a plurality of QC polynucleotides on a surface, wherein the plurality of QC polynucleotides comprises a first primer sequence; (ii) amplifying the plurality of QC polynucleotides

based on the first primer sequence; (iii) sequencing the plurality of QC polynucleotides; and (iv) aligning the plurality of QC polynucleotides against a reference to estimate an error rate in the data polynucleotides. a synthesis uniformity in the data polynucleotides, or a combination thereof for QC the data polynucleotides. In some instances, the error rate, the synthesis uniformity, or a combination thereof is based at least in part on a relative read count of the plurality of QC polynucleotides. In some instances, the plurality of QC polynucleotides is about or less 1% of the polynucleotides on the surface. In some instances, the plurality of QC polynucleotides are provided at a portion of the surface. In some instances, the plurality of QC polynucleotides are provided in uniformly on the surface. In some instances, the data polynucleotides comprise a second primer sequence. In some instances, the first primer sequence is different than the second primer sequence. In some instances, the first primer sequence and the second primer sequence are different lengths. In some instances, the quality control is performed after synthesis of the data polynucleotides. In some instances, the QC is performed prior to cleavage of the data polynucleotides from the surface. In some instances, each of the QC polynucleotides is about 50 to 200 nucleobases in length. In some instances, each of the data polynucleotides is about 100 to about 300 nucleobases in length.

[006] Further provided herein A method for quality control (QC) of data polynucleotides, comprising: (i) selecting a subset of a plurality of data polynucleotides; (ii) applying an inner codec to the subset of the plurality of data polynucleotides, wherein the inner codec comprises probabilistic decoding; and (iii) estimating an error rate in the plurality of polynucleotides based at least in part on a likelihood associated with each decoded sequence in the subset of the plurality of data polynucleotides. In some instances, a high likelihood is associated with a lower error rate. In some instances, a low likelihood is associated with a higher error rate. In some instances, further comprising decoding an index of the subset of the data polynucleotides. In some instances, the index is decoded using the inner codec, an outer codec, or a combination thereof. In some instances, the index is used to estimate a relative distribution of the subset of the plurality of data polynucleotides. In some instances, the QC is performed during synthesis of the polynucleotides, QC of stored polynucleotides, or a combination thereof. In some instances, the subset of the plurality of the data polynucleotides are selected at random. In some instances, the subset of the plurality of the data polynucleotides are selected based at least in part on their location on a surface. In some instances, the plurality of data polynucleotides comprises about 100,000 polynucleotides. In some instances, the subset of the plurality of data polynucleotides is about 0.1 % of the plurality of data polynucleotides. In some instances, the method is used in conjunction with current sensing, optical imaging, flow sensing, size estimation, quality estimation, mass estimation, or any combination thereof. In some instances, the current sensing comprises measuring a current of a chip or a section of the chip. In some instances, the current is compared to a reference value. In some instances, a difference between the current and the reference value is indicative of a chip failure, a deblocking failure, or a combination thereof. In some instances, the current sensing is performed before synthesis of the plurality of data polynucleotides. In some instances, the current sensing is used to detect a chip defect, adjust

polynucleotide synthesis locations on a chip, or a combination thereof. In some instances, mass estimation is performed using fluorescence. In some instances, the fluorescence is used to detect a yield of the plurality of polynucleotides. In some instances, optical imaging comprises detecting a chip defect, non-uniformity, or a combination thereof.

[007] Also provided herein is a method of performing QC of a plurality of cells on a surface, comprising: (i) measuring a current of each cell in the plurality of cells on the surface; (ii) determining if one or more cells in the plurality of cells comprises a defect based at least in part on the current; and (iii) synthesizing and/or storing polynucleotides at a second one or more cells in the plurality of cells, wherein the second one or more cells do not comprise the defect. In some instances, the defect comprises a physical defect. In some instances, the surface is a synthesis surface, a storage surface, or a combination thereof. In some instances, further comprising blocking the one or more cells comprising the defect. In some instances, blocking is performed by a protecting group on the surface. In some instances, blocking is performed by a photolabile protecting group on the surface. In some instances, blocking is performed by selectively supplying energy to the one or more cells. In some instances, blocking is performed by a masking material. In some instances, blocking is performed by addressable control of each cell in the plurality of cells.

BRIEF DESCRIPTION OF THE DRAWINGS

[008] A better understanding of the features and advantages of the present subject matter will be obtained by reference to the following detailed description that sets forth illustrative embodiments and the accompanying drawings of which:

[009] FIG. 1 shows a non-limiting example of quality control of polynucleotides post-synthesis in accordance with some embodiments.

[010] FIG. 2 shows a non-limiting example of periodic quality control of stored polynucleotides in accordance with some embodiments.

[011] FIG. 3 shows a non-limiting example of digital information storage in accordance with some embodiments.

[012] FIG. 4 shows a non-limiting example of generating a hash in accordance with some embodiments.

[013] FIG. 5 shows a non-limiting example of an encoding scheme, including an outer codec, in accordance with some embodiments.

[014] FIG. 6 shows a non-limiting example of an encoding scheme, including shuffling lanes of binary data, in accordance with some embodiments.

[015] FIG. 7 shows a non-limiting example of an encoding scheme, including an inner codec, in accordance with some embodiments.

[016] FIG. 8 shows a non-limiting example of an encoding scheme, including an alternative inner codec, in accordance with some embodiments.

[017] FIG. 9 shows a non-limiting example of a decoding scheme, including an inner codec and an outer codec, in accordance with some embodiments.

[018] FIG. 10 shows a non-limiting example of a greedy algorithm for decoding in accordance with some embodiments.

[019] FIG. 11 shows a non-limiting example of a maximum likelihood (ML) algorithm for decoding in accordance with some embodiments.

[020] FIG. 12 shows a non-limiting example of a computing device; in this case, a device with one or more processors, memory, storage, and a network interface.

DETAILED DESCRIPTION

[021] Provided herein are methods and systems for quality control of digital information stored in nucleic acids. For DNA data storage to be a viable option for long-lasting storage, scalable, efficient, and cost-effective methods for verifying proper synthesis and/or storage of a polynucleotide pools can be critical. As such, provided herein are quality control methods for verifying digital information encoded in nucleic acids, referred to herein as data polynucleotides. In some cases, the quality control method comprises using designated quality control (QC) polynucleotides that are synthesized and/or stored with data polynucleotides. In some cases, the quality control method comprises verifying a subset of data polynucleotides. The methods provided herein use QC polynucleotides or a subset of the data polynucleotides as a proxy to estimate an error rate, uniformity, or a combination thereof in the data polynucleotides.

[022] In some instances, the methods provide for quality control (QC) of data polynucleotides. In some instances, the method comprises providing a plurality of QC polynucleotides on a surface. In some examples, the plurality of QC polynucleotides comprises a first primer sequence. In some instances, the method comprises amplifying the plurality of QC polynucleotides based on the first primer sequence. In some instances, the method comprises sequencing the plurality of QC polynucleotides. In some instances, the method comprises aligning the plurality of QC polynucleotides against a reference. In some examples, aligning is done to estimate an error rate in the data polynucleotides, a synthesis uniformity in the data polynucleotides, or a combination thereof for QC the data polynucleotides.

[023] In some instances, the methods provide for quality control (QC) of data polynucleotides. In some instances, the method comprises selecting a subset of a plurality of data polynucleotides. In some instances, the method comprises applying an inner codec to the subset of the plurality of data polynucleotides. In some examples, the inner codec comprises probabilistic decoding. In some instances, the method comprises estimating an error rate in the plurality of polynucleotides. In some examples, the error rate in the plurality of polynucleotides is based at least in part on a likelihood associated with each

decoded sequence in the subset of the plurality of data polynucleotides.

[024] In some instances, the methods are for performing quality control (QC) of a plurality of cells on a synthesis surface. In some instances, the method comprises measuring a current of each cell in the plurality of cells on the surface. In some instances, the method comprises determining if one or more cells in the plurality of cells comprises a defect. In some examples, determining the defect is based at least in part on the current. In some instances, the method comprises synthesizing and/or storing polynucleotides at a second one or more cells in the plurality of cells. In some examples, the second one or more cells do not comprise the defect.

[025] Nucleic Acid Based Information Storage

[026] Provided herein are devices, compositions, systems and methods for nucleic acid-based information (data) storage. A biomolecule such as a DNA molecule provides a suitable host for information storage in-part due to its stability over time and capacity for enhanced information coding, as opposed to traditional binary information coding. In a first step, a digital sequence encoding an item of information (i.e., digital information in a binary code for processing by a computer) is received. An encryption scheme is applied to convert the digital sequence from a binary code to a nucleic acid sequence. A surface material for nucleic acid extension, a design for loci for nucleic acid extension (aka, arrangement spots), and reagents for nucleic acid synthesis are selected. The surface of a structure is prepared for nucleic acid synthesis. De novo polynucleotide synthesis is then performed. The synthesized polynucleotides are stored and available for subsequent release, in whole or in part. Once released, the polynucleotides, in whole or in part, are sequenced, subject to decryption to convert nucleic sequence back to digital sequence. The digital sequence is then assembled to obtain an alignment encoding for the original item of information.

[027] Items of Information

[028] Optionally, an early step of data storage process disclosed herein includes obtaining or receiving one or more items of information in the form of an initial code. Items of information (e.g., digital information) include, without limitation, text, audio and visual information. Exemplary sources for items of information include, without limitation, books, periodicals, electronic databases, medical records, letters, forms, voice recordings, animal recordings, biological profiles, broadcasts, films, short videos, emails, bookkeeping phone logs, internet activity logs, drawings, paintings, prints, photographs, pixelated graphics, and software code. Exemplary biological profile sources for items of information include, without limitation, gene libraries, genomes, gene expression data, and protein activity data. Exemplary formats for items of information include, without limitation, .txt, .PDF, .doc, .docx, .ppt, .pptx, .xls, .xlsx, .rtf, .jpg, .gif, .psd, .bmp, .tiff, .png, and .mpeg. The amount of individual file sizes encoding for an item of information, or a plurality of files encoding for items of information, in digital format include, without limitation, up to 1024 bytes (equal to 1 KB), 1024 KB (equal to 1 MB), 1024 MB (equal to 1 GB), 1024 GB (equal to 1 TB), 1024 TB (equal to 1 PB), 1 exabyte, 1 zettabyte, 1 yottabyte, 1 xenottabyte or more.

In some instances, an amount of digital information is at least 1 gigabyte (GB). In some instances, the amount of digital information is at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more than 1000 gigabytes. In some instances, the amount of digital information is at least 1 terabyte (TB). In some instances, the amount of digital information is at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more than 1000 terabytes. In some instances, the amount of digital information is at least 1 petabyte (PB). In some instances, the amount of digital information is at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more than 1000 petabytes. In some instances, the digital information does not contain genomic data acquired from an organism. Items of information in some instances are encoded. Non-limiting encoding method examples include 1 bit/base, 2 bit/base, 4 bit/base or other encoding method.

[029] Systems and Methods for Quality Control of Polynucleotide Pools

[030] Provided herein are systems and methods for QC of a polynucleotide pool or a plurality of polynucleotide pools. In some cases, the polynucleotide pool or the plurality of polynucleotide pools comprise data polynucleotides. In some instances, the data polynucleotides comprise digital information, such as binary data. In some instances, the digital information comprises an item of information, such as, but not limited to, those described herein. In some cases, the polynucleotide pool or the plurality of polynucleotide pools comprises one or more items of information. In some instances, the one or more items of information are encoded in data polynucleotides in a polynucleotide pool or a plurality of polynucleotide pools.

[031] Provided herein are systems and methods for QC of data polynucleotides. Data polynucleotides can encode digital information as described herein. In some cases, the digital information is encoded as data polynucleotides using the systems and methods described herein. However, in some instances, the QC methods described herein are agnostic to the systems and methods of encoding digital information in polynucleotides. In some instances, the QC methods described herein are agnostic to the size or type of the digital information encoded in polynucleotides. In some instances, the QC methods described herein are agnostic to the size of the polynucleotide pool described herein.

[032] In some cases, QC of data polynucleotides comprises a plurality of QC polynucleotides. The QC polynucleotides can be provided on a surface, for example, for synthesis and/or storage of the data polynucleotides, such as those described herein. In some cases, the QC polynucleotides are synthesized at the same time as the data polynucleotides. In some cases, the QC polynucleotides are provided on a surface or a portion of a surface. In some cases, the QC polynucleotides are provided uniformly on a surface or a portion of a surface. In some instances, the portion of the surface comprises a discrete location (e.g., loci, cell, feature, etc.) or a plurality of discrete locations on the surface. In some cases, the QC polynucleotides are about 1 % to about 10 % of the polynucleotides on the surface. In some instances, the polynucleotides on the surface comprise the QC polynucleotides and the data polynucleotides encoding digital information. In some cases, the QC polynucleotides are about 1 % to

about 2 %, about 1 % to about 3 %, about 1 % to about 4 %, about 1 % to about 5 %, about 1 % to about 6 %, about 1 % to about 7 %, about 1 % to about 8 %, about 1 % to about 9 %, about 1 % to about 10 %, about 2 % to about 3 %, about 2 % to about 4 %, about 2 % to about 5 %, about 2 % to about 6 %, about 2 % to about 7 %, about 2 % to about 8 %, about 2 % to about 9 %, about 2 % to about 10 %, about 3 % to about 4 %, about 3 % to about 5 %, about 3 % to about 6 %, about 3 % to about 7 %, about 3 % to about 8 %, about 3 % to about 9 %, about 3 % to about 10 %, about 4 % to about 5 %, about 4 % to about 6 %, about 4 % to about 7 %, about 4 % to about 8 %, about 4 % to about 9 %, about 4 % to about 10 %, about 5 % to about 6 %, about 5 % to about 7 %, about 5 % to about 8 %, about 5 % to about 9 %, about 5 % to about 10 %, about 6 % to about 7 %, about 6 % to about 8 %, about 6 % to about 9 %, about 6 % to about 10 %, about 7 % to about 8 %, about 7 % to about 9 %, about 7 % to about 10 %, about 8 % to about 9 %, about 8 % to about 10 %, or about 9 % to about 10 % of the polynucleotides on the surface. In some cases, the QC polynucleotides are about 1 %, about 2 %, about 3 %, about 4 %, about 5 %, about 6 %, about 7 %, about 8 %, about 9 %, or about 10 % of the polynucleotides on the surface. In some cases, the QC polynucleotides are at least about 1 %, about 2 %, about 3 %, about 4 %, about 5 %, about 6 %, about 7 %, about 8 %, or about 9 % of the polynucleotides on the surface. In some cases, the QC polynucleotides are at most about 2 %, about 3 %, about 4 %, about 5 %, about 6 %, about 7 %, about 8 %, about 9 %, or about 10 % of the polynucleotides on the surface.

[033] In some cases, the length of each of the plurality of QC polynucleotides is about 20 to about 500 bases. In some cases, the length of each of the plurality of QC polynucleotides is about 20 bases to about 50 bases, about 20 bases to about 100 bases, about 20 bases to about 200 bases, about 20 bases to about 300 bases, about 20 bases to about 400 bases, about 20 bases to about 500 bases, about 50 bases to about 100 bases, about 50 bases to about 200 bases, about 50 bases to about 300 bases, about 50 bases to about 400 bases, about 50 bases to about 500 bases, about 100 bases to about 200 bases, about 100 bases to about 300 bases, about 100 bases to about 400 bases, about 100 bases to about 500 bases, about 200 bases to about 300 bases, about 200 bases to about 400 bases, about 200 bases to about 500 bases, about 300 bases to about 400 bases, about 300 bases to about 500 bases, or about 400 bases to about 500 bases. In some cases, the length of each of the plurality of QC polynucleotides is about 20 bases, about 50 bases, about 100 bases, about 200 bases, about 300 bases, about 400 bases, or about 500 bases. In some cases, the length of each of the plurality of QC polynucleotides is at least about 20 bases, about 50 bases, about 100 bases, about 200 bases, about 300 bases, or about 400 bases. In some cases, the length of each of the plurality of QC polynucleotides is at most about 50 bases, about 100 bases, about 200 bases, about 300 bases, about 400 bases, or about 500 bases.

[034] In some cases, the plurality of QC polynucleotides each comprise a first primer sequence. In some cases, the first primer sequence of each of the plurality of QC polynucleotides is different than a second primer sequence of each of the data polynucleotides. In some instances, the first primer sequence of the plurality of QC polynucleotides is a different length than a second primer sequence of data polynucleotides. In some instances, the first primer sequence is unique to a polynucleotide pool. As an

example, a first plurality of QC polynucleotides for QC of a first polynucleotide pool encoding a first file can have a different primer sequence than a second plurality of QC polynucleotides for QC of a second polynucleotide pool encoding a second file. Alternatively, a polynucleotide pool can comprise two or more files and two or more QC polynucleotides can each have a unique primer sequence for QC of each of the files. Alternatively, a plurality of QC polynucleotides with a primer sequence can be used for QC of a plurality of polynucleotide pools. In some cases, the length of the first primer sequence is about 10 bases to about 50 bases. In some cases, the length of the first primer sequence is about 10 bases to about 15 bases, about 10 bases to about 18 bases, about 10 bases to about 20 bases, about 10 bases to about 22 bases, about 10 bases to about 25 bases, about 10 bases to about 28 bases, about 10 bases to about 30 bases, about 10 bases to about 35 bases, about 10 bases to about 40 bases, about 10 bases to about 45 bases, about 10 bases to about 50 bases, about 15 bases to about 18 bases, about 15 bases to about 20 bases, about 15 bases to about 22 bases, about 15 bases to about 25 bases, about 15 bases to about 28 bases, about 15 bases to about 30 bases, about 15 bases to about 35 bases, about 15 bases to about 40 bases, about 15 bases to about 45 bases, about 15 bases to about 50 bases, about 18 bases to about 20 bases, about 18 bases to about 22 bases, about 18 bases to about 25 bases, about 18 bases to about 28 bases, about 18 bases to about 30 bases, about 18 bases to about 35 bases, about 18 bases to about 40 bases, about 18 bases to about 45 bases, about 18 bases to about 50 bases, about 20 bases to about 22 bases, about 20 bases to about 25 bases, about 20 bases to about 28 bases, about 20 bases to about 30 bases, about 20 bases to about 35 bases, about 20 bases to about 40 bases, about 20 bases to about 45 bases, about 20 bases to about 50 bases, about 22 bases to about 25 bases, about 22 bases to about 28 bases, about 22 bases to about 30 bases, about 22 bases to about 35 bases, about 22 bases to about 40 bases, about 22 bases to about 45 bases, about 22 bases to about 50 bases, about 25 bases to about 28 bases, about 25 bases to about 30 bases, about 25 bases to about 35 bases, about 25 bases to about 40 bases, about 25 bases to about 45 bases, about 25 bases to about 50 bases, about 28 bases to about 30 bases, about 28 bases to about 35 bases, about 28 bases to about 40 bases, about 28 bases to about 45 bases, about 28 bases to about 50 bases, about 30 bases to about 35 bases, about 30 bases to about 40 bases, about 30 bases to about 45 bases, about 30 bases to about 50 bases, about 35 bases to about 40 bases, about 35 bases to about 45 bases, about 35 bases to about 50 bases, about 40 bases to about 45 bases, about 40 bases to about 50 bases, or about 45 bases to about 50 bases. In some cases, the length of the first primer sequence is about 10 bases, about 15 bases, about 18 bases, about 20 bases, about 22 bases, about 25 bases, about 28 bases, about 30 bases, about 35 bases, about 40 bases, about 45 bases, or about 50 bases. In some cases, the length of the first primer sequence is at least about 10 bases, about 15 bases, about 18 bases, about 20 bases, about 22 bases, about 25 bases, about 28 bases, about 30 bases, about 35 bases, about 40 bases, or about 45 bases. In some cases, the length of the first primer sequence is at most about 15 bases, about 18 bases, about 20 bases, about 22 bases, about 25 bases, about 28 bases, about 30 bases, about 35 bases, about 40 bases, about 45 bases, or about 50 bases.

[035] The QC polynucleotides described herein can be extracted and/or amplified. In some cases, the

plurality of QC polynucleotides are amplified. In some cases, the plurality of QC polynucleotides are amplified based on their primer sequence (e.g., first primer sequence). In some cases, the plurality of QC polynucleotides are extracted and/or amplified from surfaces where they are synthesized or stored. After extraction and/or amplification of QC polynucleotides from the surface of a structure, suitable sequencing technology may be employed to sequence the polynucleotides, as further described herein. In some cases, the DNA sequence is read on the substrate or within a feature of a structure.

[036] In some cases, the plurality of QC polynucleotides are aligned. In some cases, the plurality of QC polynucleotides are aligned against a reference. In some instances, the reference is a known sequence or a preselected sequence. In some instances, the known sequence or the preselected sequence is the original sequence of the plurality of QC polynucleotides. In some cases, the plurality of QC polynucleotides are aligned against a reference to estimate an error rate in the data polynucleotides, a synthesis uniformity in the data polynucleotides, or a combination thereof.

[037] An error rate in the data polynucleotides can be estimated by aligning the plurality of QC polynucleotides against a reference. In some cases, the reference is a known sequence, as previously described herein. In some cases, aligning the plurality of QC polynucleotides against a reference generates a relative read count. In some instances, the relative read count comprises the number of sequence QC polynucleotides that have the same sequence as the reference sequence. In some instances, the relative read count is used to estimate an error rate in the plurality of QC polynucleotides by determining the number of sequenced QC polynucleotides that have the same sequence as the reference sequence out of all the sequenced QC polynucleotides. In some instances, the error rate in the plurality of QC polynucleotides is used to estimate an error rate in the data polynucleotides. In some instances, the error rate in the data polynucleotides is based at least in part on the relative read count.

[038] A synthesis uniformity in the data polynucleotides can be estimated by aligning the plurality of QC polynucleotides against a reference. In some cases, the reference is a known sequence, as previously described herein. In some cases, aligning the plurality of QC polynucleotides against a reference generates a relative read count, as described herein. In some instances, the relative read count is used to estimate a synthesis uniformity in the plurality of QC polynucleotides by determining the number of sequenced QC polynucleotides that have the same sequence as the reference sequence out of all the sequenced QC polynucleotides. In some instances, the relative read count is used to estimate a synthesis uniformity in one or more discrete locations (e.g., loci, cell, feature, etc.) where the sequenced QC polynucleotides have the same sequence as the reference sequence. For example, it may be determined that a particular cell out of a plurality of cells comprises less QC polynucleotides or comprises QC polynucleotides that comprise less alignment with the reference compared to QC polynucleotides in other cells. In some instances, the synthesis uniformity in the plurality of QC polynucleotides is used to estimate a synthesis uniformity in the data polynucleotides. In some instances, the synthesis uniformity in the data polynucleotides is based at least in part on the relative read count.

[039] In some cases, the methods for QC of polynucleotides comprising QC polynucleotides, as described herein, are performed after synthesis of the data polynucleotides. In some cases, the methods for QC of polynucleotides comprising QC polynucleotides, as described herein, are performed after an initial synthesis of the data polynucleotides, as exemplary illustrated in FIG. 1. In some cases, the methods for QC of polynucleotides comprising QC polynucleotides, as described herein, are performed after re-synthesis of the data polynucleotides if a synthesis or storage error is encountered. In some cases, the methods for QC of polynucleotides comprising QC polynucleotides, as described herein, are performed on data polynucleotides that are stored.

[040] In some cases, QC of data polynucleotides comprises selecting a subset of a plurality of data polynucleotides. In some instances, the subset of the plurality of data polynucleotides is selected randomly. In some instances, the subset of the plurality of data polynucleotides is selected pseudo randomly. In some instances, the subset of the plurality of data polynucleotides are selected at least in part based on their location on a synthesis or storage surface, such as those described herein. In some instances, the subset of the plurality of data polynucleotides are selected at least in part based on one or more physical or chemical properties, such as, but not limited to, those measured by current sensing, optical imaging, flow sensing, etc. In some cases, the subset of the plurality of data polynucleotides comprises about 0.01 % to about 5 % of the plurality of data polynucleotides. In some cases, the subset of the plurality of data polynucleotides comprises about 0.01 % to about 0.02 %, about 0.01 % to about 0.05 %, about 0.01 % to about 0.08 %, about 0.01 % to about 0.1 %, about 0.01 % to about 0.2 %, about 0.01 % to about 0.5 %, about 0.01 % to about 1 %, about 0.01 % to about 2 %, about 0.01 % to about 3 %, about 0.01 % to about 4 %, about 0.01 % to about 5 %, about 0.02 % to about 0.05 %, about 0.02 % to about 0.08 %, about 0.02 % to about 0.1 %, about 0.02 % to about 0.2 %, about 0.02 % to about 0.5 %, about 0.02 % to about 1 %, about 0.02 % to about 2 %, about 0.02 % to about 3 %, about 0.02 % to about 4 %, about 0.02 % to about 5 %, about 0.05 % to about 0.08 %, about 0.05 % to about 0.1 %, about 0.05 % to about 0.2 %, about 0.05 % to about 0.5 %, about 0.05 % to about 1 %, about 0.05 % to about 2 %, about 0.05 % to about 3 %, about 0.05 % to about 4 %, about 0.05 % to about 5 %, about 0.08 % to about 0.1 %, about 0.08 % to about 0.2 %, about 0.08 % to about 0.5 %, about 0.08 % to about 1 %, about 0.08 % to about 2 %, about 0.08 % to about 3 %, about 0.08 % to about 4 %, about 0.08 % to about 5 %, about 0.1 % to about 0.2 %, about 0.1 % to about 0.5 %, about 0.1 % to about 1 %, about 0.1 % to about 2 %, about 0.1 % to about 3 %, about 0.1 % to about 4 %, about 0.1 % to about 5 %, about 0.2 % to about 0.5 %, about 0.2 % to about 1 %, about 0.2 % to about 2 %, about 0.2 % to about 3 %, about 0.2 % to about 4 %, about 0.2 % to about 5 %, about 0.5 % to about 1 %, about 0.5 % to about 2 %, about 0.5 % to about 3 %, about 0.5 % to about 4 %, about 0.5 % to about 5 %, about 1 % to about 2 %, about 1 % to about 3 %, about 1 % to about 4 %, about 1 % to about 5 %, about 2 % to about 3 %, about 2 % to about 4 %, about 2 % to about 5 %, about 3 % to about 4 %, about 3 % to about 5 %, or about 4 % to about 5 % of the plurality of data polynucleotides. In some cases, the subset of the plurality of data polynucleotides comprises about 0.01 %, about 0.02 %, about 0.05 %, about 0.08 %, about 0.1 %, about 0.2 %, about 0.5

%, about 1 %, about 2 %, about 3 %, about 4 %, or about 5 % of the plurality of data polynucleotides. In some cases, the subset of the plurality of data polynucleotides comprises at least about 0.01 %, about 0.02 %, about 0.05 %, about 0.08 %, about 0.1 %, about 0.2 %, about 0.5 %, about 1 %, about 2 %, about 3 %, or about 4 % of the plurality of data polynucleotides. In some cases, the subset of the plurality of data polynucleotides comprises at most about 0.02 %, about 0.05 %, about 0.08 %, about 0.1 %, about 0.2 %, about 0.5 %, about 1 %, about 2 %, about 3 %, about 4 %, or about 5 % of the plurality of data polynucleotides.

[041] In some cases, a polynucleotide pool comprises the plurality of data polynucleotides. In some cases, the plurality of data polynucleotides comprise about 100 to 500,000 polynucleotides. In some cases, the plurality of data polynucleotides comprise about 100 to about 500, about 100 to about 1,000, about 100 to about 5,000, about 100 to about 10,000, about 100 to about 50,000, about 100 to about 100,000, about 100 to about 200,000, about 100 to about 300,000, about 100 to about 400,000, about 100 to about 500,000, about 500 to about 1,000, about 500 to about 5,000, about 500 to about 10,000, about 500 to about 50,000, about 500 to about 100,000, about 500 to about 200,000, about 500 to about 300,000, about 500 to about 400,000, about 500 to about 500,000, about 1,000 to about 5,000, about 1,000 to about 10,000, about 1,000 to about 50,000, about 1,000 to about 100,000, about 1,000 to about 200,000, about 1,000 to about 300,000, about 1,000 to about 400,000, about 1,000 to about 500,000, about 5,000 to about 10,000, about 5,000 to about 50,000, about 5,000 to about 100,000, about 5,000 to about 200,000, about 5,000 to about 300,000, about 5,000 to about 400,000, about 5,000 to about 500,000, about 10,000 to about 50,000, about 10,000 to about 100,000, about 10,000 to about 200,000, about 10,000 to about 300,000, about 10,000 to about 400,000, about 10,000 to about 500,000, about 50,000 to about 100,000, about 50,000 to about 200,000, about 50,000 to about 300,000, about 50,000 to about 400,000, about 50,000 to about 500,000, about 100,000 to about 200,000, about 100,000 to about 300,000, about 100,000 to about 400,000, about 100,000 to about 500,000, about 200,000 to about 300,000, about 200,000 to about 400,000, about 200,000 to about 500,000, about 300,000 to about 400,000, about 300,000 to about 500,000, or about 400,000 to about 500,000 polynucleotides. In some cases, the plurality of data polynucleotides comprise about 100, about 500, about 1,000, about 5,000, about 10,000, about 50,000, about 100,000, about 200,000, about 300,000, about 400,000, or about 500,000 polynucleotides. In some cases, the plurality of data polynucleotides comprise at least about 100, about 500, about 1,000, about 5,000, about 10,000, about 50,000, about 100,000, about 200,000, about 300,000, or about 400,000 polynucleotides. In some cases, the plurality of data polynucleotides comprise at most about 500, about 1,000, about 5,000, about 10,000, about 50,000, about 100,000, about 200,000, about 300,000, about 400,000, or about 500,000 polynucleotides.

[042] In some cases, an inner codec is applied to the subset of the plurality of data polynucleotides. In some instances, the inner codec comprises probabilistic decoding. An inner codec generally comprises a decoding polynucleotides into digital information. In some cases, the inner codec comprises converting or transforming each of the subset of the plurality of data polynucleotides into binary data. In some

instances, a full length of the subset of the plurality of data polynucleotides are transformed or converted into binary data (e.g., full decoding). In some instances, a partial length of the subset of the plurality of data polynucleotides are transformed or converted into binary data (e.g., partial decoding). In some examples, the partial length comprises an index, such as those described herein (e.g., lane index, frame index, UUID, content ID, etc.). In some instances, the inner codec is applied to the subset of the plurality of data polynucleotides that have been sequenced. In some instances, the inner codec is applied to the subset of the plurality of data polynucleotides that have or have not been ordered, aligned, clustered, or any combination thereof.

[043] In some cases, the plurality of data polynucleotides and/or the subset of the plurality of data polynucleotides are encoded using the methods described herein. In some cases, the plurality of data polynucleotides and/or the subset of the plurality of data polynucleotides are decoded using the methods described herein. In some instances, the inner codec comprises a greedy algorithm. In some instances, the inner codec comprises a maximum likelihood (ML) algorithm. In some instances, the inner codec comprises a mixed greedy ML algorithm.

[044] In some cases, the probabilistic decoding of the inner codec provides a likelihood of the overall decoded sequence. In some instances, redundancy within each polynucleotide sequences helps to estimate error rates without knowing a reference polynucleotide. As an example, if the inner codec decodes sequences with high probabilities and/or in very few steps, then the error rate is likely low. As a further example, if the inner codec decodes sequences with low probabilities and/or takes more steps, then the error rate is likely high.

[045] In some cases, the data polynucleotides comprises an index. In some instances, an index of the subset of the plurality of data polynucleotides is decoded. In some instances, the index is decoded using an inner codec, an outer codec, or a combination thereof, such as, but not limited to, those described herein. In some instances, the index is used to estimate a relative distribution of the subset of the plurality of polynucleotides. In some examples, the relative distribution is used to estimate uniformity of the data polynucleotides. For example, if the plurality of data polynucleotides comprises about 100,000 polynucleotide sequences, and the subset selected is 0.1% of the data polynucleotides, a distribution centered around 100 decoded indexes can be expected. In some examples, relative distribution changes between subsets of the data polynucleotides indicate a loss of uniformity across the data polynucleotides.

[046] In some cases, the methods for QC of data polynucleotides comprising selecting a subset, as described herein, are performed after synthesis of the data polynucleotides. In some cases, the methods for QC of polynucleotides comprising selecting a subset, as described herein, are performed after an initial synthesis of the data polynucleotides, as exemplary illustrated in **FIG. 1**. In some cases, the methods for QC of polynucleotides comprising selecting a subset, as described herein, are performed after re-synthesis of the data polynucleotides if an synthesis or storage error is encountered. In some cases, the methods for QC of polynucleotides comprising selecting a subset, as described herein, are

performed on data polynucleotides that are stored, as exemplary illustrated in **FIG. 2**.

[047] In some cases, the methods and systems for QC of polynucleotides are used in conjunction with one or more additional methods for QC. In some instances, the one or more additional methods comprise any one of: current sensing, resistance sensing, optical imaging, flow sensing, size estimation, quality estimation, mass estimation, or any combination thereof. In some cases, current or resistance sensing comprises measuring a current or a resistance, respectively, of a chip or a section of the chip. In some instances, a current or resistance is compared to a reference value. In some instances, the difference between the current or resistance and the standard value is indicative of a chip failure, synthesis error, or a combination thereof. In some examples, the chip failure comprises a defect in a chip. In some instances, the defect in the chip causes a polynucleotide synthesis and/or storage problem. In some examples, the synthesis error comprises a deblocking failure. In some cases, mass estimation comprises measuring absorbance to estimate a mass of a polynucleotide sequence. In some cases, mass estimation comprises using fluorescence to measure the mass of a polynucleotide sequence. In some instances, the fluorescence is used to detect a yield of the plurality of polynucleotides. In some cases, optical imaging comprises detecting a chip defect, non-uniformity, or a combination thereof. In some cases, flow sensing is used to detect the flow of a liquid, gas, or a combination thereof over the synthesis and/or storage chip.

[048] An exemplary flow diagram of QC of a polynucleotide pool is provided in **FIG. 1**. As shown, current sensing may be employed prior to synthesis to QC a synthesis chip. Current sensing prior to synthesis can be performed to detect a chip defect, adjust polynucleotide synthesis locations on the chip, or a combination thereof. This can be followed by synthesis placement optimizations, followed by the synthesis of data polynucleotides. In some cases, the data polynucleotides are synthesized along with QC polynucleotides, as previously described herein. In some cases, synthesis of the plurality of data polynucleotides is performed with continuous QC using one or more additional methods described herein. In some instances, the continuous QC comprises current sensing, resistance sensing, optical imaging, flow sensing, or a combination thereof. In some cases, post-synthesis QC comprises determining an oligo length distribution and/or mass estimation of the plurality of data polynucleotides using, for example, techniques described herein.

[049] In some cases, the QC polynucleotides are amplified and sequenced for QC of the data polynucleotides, as previously described herein. In some instances, the QC polynucleotides are fully sequenced. In some instances, the QC polynucleotides are partially sequenced. In some cases, the QC polynucleotides are aligned, and an error rate and/or a uniformity is estimated, as previously described herein.

[050] Alternatively or in combination, in some cases, the data polynucleotides are amplified and a subset (or sub-sample) is sequenced. In some cases, the subset is fully decoded. In some cases, the subset is partially decoded. In some instances, partial decoding comprises applying an inner codec, an outer codec, or a combination thereof. In some examples, an inner codec is applied to estimate an error rate, as

previously described herein. In some instances, an index of the subset is partially decoded. In some examples, an outer codec is applied to estimate a uniformity, as previously described herein.

[051] In some cases, QC of the QC polynucleotides, data polynucleotides, or a combination thereof is used to determine a final QC decision. In some instances, the final QC decision is based on the error rate, uniformity, or both. In some instances, the final QC decision comprises a pass or fail of the synthesized data polynucleotides. In some cases, if the final QC decision is a pass, then the data polynucleotides are stored. In some cases, if the final QC decision is a fail, then the data polynucleotides are resynthesized. In some cases, the final QC decision comprises a pass for some sections of a synthesis surface and a fail for some sections of the synthesis surface. In some instances, only the data polynucleotides from the pass sections of the synthesis surface are stored. In some instances the data polynucleotides from the fail sections of the synthesis surface are resynthesized.

[052] In some cases, the final QC decision comprises a threshold. In some instances, the threshold comprises a static value or a dynamic value. In some instances, the threshold comprises a static range or a dynamic range. In some instances, the threshold is based on one or more combined values or ranges, such as error rate, uniformity, or both. In some instances, if the error rate is low and the uniformity is high, then the final QC decision is a pass. For example, the error rate may be less than 5 %, 4 %, 3 %, 2 %, 1 %, 0.5 %, 0.1 %, 0.05 %, 0.01 %, 0.0005 %, or 0.0001 %. In a further example, the uniformity may be greater than 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 %, 99.5 %, 99.9 %, 99.95 %, or 99.99 %. In some instances, if the error rate is high and the uniformity is low, then the final QC decision is a pass. For example, the error rate may be greater than 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, or 50 %. In a further example, the uniformity may be less than 80 %, 70 %, 60 %, 50 %, 40 %, 30 %, 20 %, or 10 %.

[053] A further exemplary flow diagram of QC of a polynucleotide pool is provided in **FIG. 2**. Here, a plurality of data polynucleotides in a pool are already synthesized and stored. In some cases, QC of data polynucleotides in a pool are periodically performed. In some cases, QC of data polynucleotides are performed over days, weeks, months, or years. In some cases, a plurality of data polynucleotides are retrieved from storage. In some cases, the data polynucleotides are amplified, and a sub-set (or sub-sample) is sequenced. In some cases, the subset is fully decoded. In some cases, the subset is partially decoded. In some instances, partial decoding comprises applying an inner codec, an outer codec, or a combination thereof. In some examples, an inner codec is applied to estimate an error rate, as previously described herein. In some instances, an index of the subset is partially decoded. In some examples, an outer codec is applied to estimate a uniformity, as previously described herein.

[054] In some cases, QC of the retrieved data polynucleotides further comprises a pool level QC comprising determining an oligo length distribution and/or mass estimation of the plurality of data polynucleotides using, for example, techniques described herein. In some cases, the pool level QC is combined with the sub-set QC to determine a final QC decision, as shown in **FIG. 2**. In some instances,

the final QC decision comprises a pass or fail of the synthesized data polynucleotides. In some cases, if the final QC decision is a pass, then the data polynucleotides are returned to storage. In some cases, if the final QC decision is a fail, then the data polynucleotides are fully sequenced and/or decoded. In some instances, the full sequencing and decoding of the data polynucleotides comprises sequencing and/or decoding duplicate data polynucleotides if a sample of data polynucleotides cannot be decoded. In some cases, the data polynucleotides are resynthesized. In some cases, the final QC decision comprises a pass for some sections of a storage surface and a fail for some sections of the storage surface. In some instances, only the data polynucleotides from the pass sections of the storage surface are returned to storage. In some instances, the data polynucleotides from the fail sections of the storage surface are fully sequenced, decoded, and/or resynthesized.

[055] In some cases, the final QC decision comprises a threshold. In some instances, the threshold comprises a static value or a dynamic value. In some instances, the threshold comprises a static range or a dynamic range. In some instances, the threshold is based on one or more combined values or ranges, such as error rate, uniformity, or both. In some instances, if the error rate is low and the uniformity is high, then the final QC decision is a pass. For example, the error rate may be less than 5 %, 4 %, 3 %, 2 %, 1 %, 0.5 %, 0.1 %, 0.05 %, 0.01 %, 0.0005 %, or 0.0001 %. In a further example, the uniformity may be greater than 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 %, 99.5 %, 99.9 %, 99.95 %, or 99.99 %. In some instances, if the error rate is high and the uniformity is low, then the final QC decision is a pass. For example, the error rate may be greater than 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, or 50 %. In a further example, the uniformity may be less than 80 %, 70 %, 60 %, 50 %, 40 %, 30 %, 20 %, or 10 %.

[056] Further provided herein are systems and methods for performing quality control (QC) of a plurality of cells on a surface, such as, a synthesis surface or a storage surface. In some cases, the cells comprise active regions on a surface, such as compartments, location, loci, features, spots, or any variation thereof suitable for synthesis and/or storage of polynucleotides. In some cases, the surface comprises a synthesis and/or storage surface of a device, such as, but not limited to, those described herein.

[057] In some cases, a method for performing QC of a plurality of cells comprises one or more steps. In some cases, a step comprises measuring a physical and/or chemical property of each of the plurality of cells on the surface. In some instances, a step comprises current sensing, resistance sensing, optical imaging, or a combination thereof. In some instances, a voltage is applied to the surface. In some instances, the voltage is about 0.1 V to about 3 V. In some instances, the voltage is about 0.1 V to about 0.25 V, about 0.1 V to about 0.5 V, about 0.1 V to about 0.75 V, about 0.1 V to about 1 V, about 0.1 V to about 1.25 V, about 0.1 V to about 1.5 V, about 0.1 V to about 1.75 V, about 0.1 V to about 2 V, about 0.1 V to about 2.5 V, about 0.1 V to about 3 V, about 0.25 V to about 0.5 V, about 0.25 V to about 0.75 V, about 0.25 V to about 1 V, about 0.25 V to about 1.25 V, about 0.25 V to about 1.5 V, about 0.25 V to

about 1.75 V, about 0.25 V to about 2 V, about 0.25 V to about 2.5 V, about 0.25 V to about 3 V, about 0.5 V to about 0.75 V, about 0.5 V to about 1 V, about 0.5 V to about 1.25 V, about 0.5 V to about 1.5 V, about 0.5 V to about 1.75 V, about 0.5 V to about 2 V, about 0.5 V to about 2.5 V, about 0.5 V to about 3 V, about 0.75 V to about 1 V, about 0.75 V to about 1.25 V, about 0.75 V to about 1.5 V, about 0.75 V to about 1.75 V, about 0.75 V to about 2 V, about 0.75 V to about 2.5 V, about 0.75 V to about 3 V, about 1 V to about 1.25 V, about 1 V to about 1.5 V, about 1 V to about 1.75 V, about 1 V to about 2 V, about 1 V to about 2.5 V, about 1 V to about 3 V, about 1.25 V to about 1.5 V, about 1.25 V to about 1.75 V, about 1.25 V to about 2 V, about 1.25 V to about 2.5 V, about 1.25 V to about 3 V, about 1.5 V to about 1.75 V, about 1.5 V to about 2 V, about 1.5 V to about 2.5 V, about 1.5 V to about 3 V, about 1.75 V to about 2 V, about 1.75 V to about 2.5 V, about 1.75 V to about 3 V, about 2 V to about 2.5 V, about 2 V to about 3 V, or about 2.5 V to about 3 V. In some instances, the voltage is about 0.1 V, about 0.25 V, about 0.5 V, about 0.75 V, about 1 V, about 1.25 V, about 1.5 V, about 1.75 V, about 2 V, about 2.5 V, or about 3 V. In some instances, the voltage is at least about 0.1 V, about 0.25 V, about 0.5 V, about 0.75 V, about 1 V, about 1.25 V, about 1.5 V, about 1.75 V, about 2 V, or about 2.5 V. In some instances, the voltage is at most about 0.25 V, about 0.5 V, about 0.75 V, about 1 V, about 1.25 V, about 1.5 V, about 1.75 V, about 2 V, about 2.5 V, or about 3 V. In some instances, a step comprises measuring a current, resistance, or a combination thereof of each of the plurality of cell on the surface when a voltage is applied.

[058] In some cases, the current sensing, resistance sensing, optical imaging, or a combination thereof is used to determine if one or more cells in the plurality of cells comprises a defect. In some instances, the defect comprises a physical defect in the surface. In some instances, the defect is determined based at least in part on the current, resistance, or a combination thereof. In some instances, a current, resistance, or a combination thereof measured in a cell comprising a defect is different than a corresponding standard value measured in cells that do not comprise a defect. In some instances, the current, resistance, or a combination thereof measured in a cell comprising a defect is different than the corresponding standard value by about 1 % to about 40 %. In some instances, the current, resistance, or a combination thereof is different by about 1 % to about 5 %, about 1 % to about 10 %, about 1 % to about 15 %, about 1 % to about 20 %, about 1 % to about 25 %, about 1 % to about 30 %, about 1 % to about 35 %, about 1 % to about 40 %, about 5 % to about 10 %, about 5 % to about 15 %, about 5 % to about 20 %, about 5 % to about 25 %, about 5 % to about 30 %, about 5 % to about 35 %, about 5 % to about 40 %, about 10 % to about 15 %, about 10 % to about 20 %, about 10 % to about 25 %, about 10 % to about 30 %, about 10 % to about 35 %, about 10 % to about 40 %, about 15 % to about 20 %, about 15 % to about 25 %, about 15 % to about 30 %, about 15 % to about 35 %, about 15 % to about 40 %, about 20 % to about 25 %, about 20 % to about 30 %, about 20 % to about 35 %, about 20 % to about 40 %, about 25 % to about 30 %, about 25 % to about 35 %, about 25 % to about 40 %, about 30 % to about 35 %, about 30 % to about 40 %, or about 35 % to about 40 %. In some instances, the current, resistance, or a combination thereof is different by about 1 %, about 5 %, about 10 %, about 15 %, about 20 %, about 25 %, about 30 %, about

35 %, or about 40 %. In some instances, the current, resistance, or a combination thereof is different by at least about 1 %, about 5 %, about 10 %, about 15 %, about 20 %, about 25 %, about 30 %, or about 35 %. In some instances, the current, resistance, or a combination thereof is different by at most about 5 %, about 10 %, about 15 %, about 20 %, about 25 %, about 30 %, about 35 %, or about 40 %.

[059] In some cases, if a defect is detected, the one or more cells comprising the defect is blocked. In some instances, blocking one or more cells comprises leaving a protecting group, such as DMT, on the surface. In some instances, blocking one or more cells comprises preventing deblocking of the protecting group. In some instances, blocking comprises one or more photolabile protecting groups, where the hydroxyl groups generated on the surface are blocked by photolabile-protecting groups. For example, when the surface is exposed to UV light, such as through a photolithographic mask, a pattern of free hydroxyl groups on the surface may be generated. These hydroxyl groups can react with photoprotected nucleoside phosphoramidites, according to phosphoramidite chemistry. In some examples, a second photolithographic mask can be applied and the surface can be exposed to UV light to generate second pattern of hydroxyl groups, followed by coupling with 5'- photoprotected nucleoside phosphoramidite. Likewise, patterns can be generated and oligomer chains can be extended. Without being bound by theory, the lability of a photocleavable group depends on the wavelength and polarity of a solvent employed and the rate of photocleavage may be affected by the duration of exposure and the intensity of light. This method can leverage a number of factors such as accuracy in alignment of the masks, efficiency of removal of photo- protecting groups, and the yields of the phosphoramidite coupling step.

[060] In some instances, blocking further comprises selectively supplying energy to one or more cells. In some instances, a mask is created on a surface through heating elements on or proximal to the surface. In some instances, a layer of masking material is applied to the surface and the heating elements are employed to apply energy to the masking material at selected sites, whereby the applied energy brings about a phase change in the masking material at the selected sites such that it adheres to the surface or can be displaced from the surface to mask or unmask the selected sites respectively. In some instances, the masking material is a solid, gas, liquid, or a combination thereof. In some instances, the masking material comprises, for example, C₁₅-C₃₀ n-alkanes (e.g., tetracosane (C₂₄), icosane (C₂₀), etc.). In some instances, the masking material comprises a mixture of two or more higher straight chain alkanes, such as, for example, C₁₆-C₃₀ n-alkanes or C₁₈-C₂₈ n-alkanes. In some instances, the masking material, for example nanospheres, may be deposited on the surface in the form of a dispersion, for example in acetonitrile.

[061] In some instances, blocking further comprises addressable locations on a surface. In some instances, the locations are addressable through one or more electrodes near the locations. In some instances, the one or more electrodes are independently addressable. In some instances, the one or more electrodes at each of the locations on a surface are independently addressable. In some instances, each electrode controls nucleoside (nucleoside phosphoramidite) coupling through electrochemistry at a

specific location on the surface. In some instances, reagents comprise protons or other acid molecule. In some instances, electrodes are located at positions around the edges of a surface of a well. In some instances, electrodes control chemical reactions occurring near the synthesis surface. For example, if acid or other reagent is generated near the synthesis surface, the portion of a polynucleotide bound to this surface will be contacted with a higher concentration of acid than the portion of the polynucleotide that is distal to the site of acid generation. This may lead to degradation of the portion of the polynucleotide which is exposed to higher concentrations of acid. Electrodes, such as those located near the surface of a well, in some instances produce or control a proton gradient which results in uniform or targeted exposure of a portion of the polynucleotide to acid. Sites near uncharged electrodes do not couple with nucleosides deposited over the synthesis surface, and the pattern of charged electrodes is altered before addition of the next nucleoside. By applying a series of electrode-controlled masks to the surface, the desired polynucleotides are synthesized at exact locations on the surface.

[062] In some cases, polynucleotides are synthesized and/or stored at a second one or more cells in the plurality of cells. In some instances, the second one or more cells do not comprise a defect. In some instances, the polynucleotides are synthesized and/or stored using the systems and methods described herein. In some instances, the polynucleotides are encoded according to the systems and methods described herein.

[063] Systems and Methods for Digital Information Storage

[064] Provided herein are methods and systems for storage of digital information. In some cases, the digital information comprises one or more objects. In some cases, the one or more objects comprises an item of information, such as, but not limited to, those described herein. In some cases, the one or more objects comprises a file or a metadata associated the file. In some cases, the digital information comprises binary data. In some cases, the binary data is a byte stream or a byte array. In some cases, each of the one or more objects is about 1 GB to about 1 TB. In some cases, the each of the one or more objects is about 1 GB to about 1 TB. In some cases, the each of the one or more objects is about 1 GB to about 10 GB, about 1 GB to about 50 GB, about 1 GB to about 100 GB, about 1 GB to about 500 GB, about 1 GB to about 1 TB, about 10 GB to about 50 GB, about 10 GB to about 100 GB, about 10 GB to about 500 GB, about 10 GB to about 1 TB, about 50 GB to about 100 GB, about 50 GB to about 500 GB, about 50 GB to about 1 TB, about 100 GB to about 500 GB, about 100 GB to about 1 TB, or about 500 GB to about 1 TB. In some cases, each of the one or more objects is about 1 GB, about 10 GB, about 50 GB, about 100 GB, about 500 GB, or about 1 TB. In some cases, each of the one or more objects is at least about 1 GB, about 10 GB, about 50 GB, about 100 GB, or about 500 GB. In some cases, each of the one or more objects is at most about 10 GB, about 50 GB, about 100 GB, about 500 GB, or about 1 TB.

[065] A system of storing digital information can comprise one or more processing units, a memory in communication with the one or more processing units, instructions stored in the memory and executed on the one or more processing units, or any combination thereof. In some cases, the one or more processing

units and memory are distributed across one or more physical or logical locations. In some cases, the one or more processing units include any combination of central processing units (CPUs), graphical processing units (GPUs), single core processors, multi-core processors, processor clusters, application-specific integrated circuits (ASICs), programmable circuits such as Field Programmable Gate Arrays (FPGA), an AI-accelerator and variations thereof. In some cases, the one or more of the processing units comprise a Single Instruction Multiple Data (SIMD) or Single Program Multiple Data (SPMD) parallel architectures. As an example, the one or more processing units include one or more GPUs or CPUs that implement SIMD or SPMD. In some instances, an AI-accelerator comprise Google-TPU, Graphcore, Cerebras, SambaNova, or a combination thereof. In some embodiments, one or more of the processing units is implemented in software and/or firmware, in addition to hardware implementations. Software or firmware implementations of the processing units can include computer- or machine- executable instructions written in any suitable programming language to perform the various functions described herein. Software implementations of the one or more processing units can be stored in whole or part in the memory. Alternatively or additionally, the system can comprise one or more hardware logic components. For example, and without limitation, illustrative types of hardware logic components that can be used include Field-programmable Gate Arrays (FPGAs), Application-specific Integrated Circuits (ASICs), Application-specific Standard Products (ASSPs), System-on-a-chip systems (SOCs), Complex Programmable Logic Devices (CPLDs), etc. In some cases, the memory comprises removable storage, non-removable storage, local storage, and/or remote storage to provide storage of instructions, data structures, program modules (e.g., hashing module), and any other data described herein. In some instances, the memory is used to store information related to the algorithms described herein (e.g., software code, parameters, executable instructions, etc.).

[066] The instructions stored on the memory can comprise one or more steps for storing digital information. In some cases, the one or more steps comprises splitting digital information of one or more objects into a plurality of pools. In some cases, each of the plurality of pools is about 1 GB to about 1 TB. In some cases, each of the plurality of pools is about 1 GB to about 1 TB. In some cases, each of the plurality of pools is about 1 GB to about 10 GB, about 1 GB to about 50 GB, about 1 GB to about 100 GB, about 1 GB to about 500 GB, about 1 GB to about 1 TB, about 10 GB to about 50 GB, about 10 GB to about 100 GB, about 10 GB to about 500 GB, about 10 GB to about 1 TB, about 50 GB to about 100 GB, about 50 GB to about 500 GB, about 50 GB to about 1 TB, about 100 GB to about 500 GB, about 100 GB to about 1 TB, or about 500 GB to about 1 TB. In some cases, each of the plurality of pools is about 1 GB, about 10 GB, about 50 GB, about 100 GB, about 500 GB, or about 1 TB. In some cases, each of the plurality of pools is at least about 1 GB, about 10 GB, about 50 GB, about 100 GB, or about 500 GB. In some cases, each of the plurality of pools is at most about 10 GB, about 50 GB, about 100 GB, about 500 GB, or about 1 TB.

[067] In some instances, the one or more objects comprises an item of information, such as a file, as previously described herein. In some instances, the one or more objects comprises a metadata associated

with an item of information (e.g., metadata associated with a file). Non-limiting examples of metadata associated with an object include a list of keywords attached to an object, an object size, a thumbnail picture, a text summary, an ID range for a sorted key-value database, a timestamp, a version, or any other data providing information about one or more aspects of an object, or any combination thereof. In some examples, the metadata is customizable. In some examples, the metadata is used to search for an object in the plurality of pools.

[068] An exemplary diagram of digital information storage is illustrated in **FIG. 3**. As shown, one or more objects **305** can be split into a plurality of pools **310**. In some cases, one object is split into a plurality of pools. In some cases, one object is split into two, three, four, five, six, seven, eight, nine, or ten pools. In some cases, more than one object is split into a plurality of pools. In some cases, one or more objects is in a pool. In some cases, one, two, three, four, five, six, seven, eight, nine, or ten objects are in a pool. In some cases, the plurality of pools are duplicated. In some cases, the plurality of pools comprise redundant pools, where two or more pools comprise the same one or more objects. In some cases, two, three, four, five, six, seven, eight, nine, or ten pools comprise the same one or more objects.

[069] Each pool in the plurality of pools can comprise any one of or a combination of a pool descriptor, a pool item, or an end descriptor. In some cases, a pool comprises at least one pool item. In some cases, a pool comprises more than one pool item. In some cases, a pool comprises at least one pool descriptor. In some cases, a pool comprises more than one pool descriptor. In some cases, a pool comprises at least one end descriptor. In some cases, a pool comprises more than one end descriptor. As an example, each pool comprises a pool descriptor **315**, one or more pool items **320**, and an end descriptor **325**. In some cases, a pool comprises redundant pool items, pool descriptors, end pool descriptors, or a combination thereof. In such cases, two or more pool items, pool descriptors, end pool descriptors, or a combination thereof are identical. In some instances, two, three, four, five, six, seven, eight, nine, or ten, pool descriptors, end pool descriptors, or a combination thereof are identical.

[070] In some cases, the one or more steps in the instructions comprise generating a pool descriptor, a pool item, an end descriptor, or any combination thereof in each pool of the plurality of pools. In some cases, the pool descriptor comprises a version, a pool ID, a list of pool item descriptors, or any combination thereof. In some instances, the version comprises the version of information (e.g., if information is updated). In some instances, the pool ID comprises a unique ID of the pool. In some examples, the unique ID comprises a universal unique identifier (UUID). In some examples, the unique ID comprises a content ID. In some examples, the content ID comprises a digital fingerprinting system, which can be used to identify and/or manage copyright or ownership of a content. In some instances, the list of pool item descriptors comprises a path of an object, a size of an object (e.g., a total size of an object), a range of the pool item within an object, offset of the pool item in a pool, or any combination thereof. In some examples, the range of the pool item within an object comprises one or more locations of a payload in the pool item within an object. In some examples, the one or more locations comprises a

start and/or an end range of a payload in a pool item (e.g., line 1-6 in pool item 1, line 7-13 in pool item 2, ... etc., in a pool). In some examples, the offset of the pool item comprises a payload location within the one or more pool items in a pool. In some cases, the pool item comprises a data payload and/or a hash of the pool item. In some instances, the data payload comprises the object or a portion of the object that is being stored. In some instances, the hash of the pool item comprises a hashed value of the object or a portion of the object that is being stored. In some cases, the end pool descriptor comprises a list of object descriptors. In some instances, the list of object descriptors comprises a path of the object and/or a hash of the object. In some examples, the path of the object comprises a unique path. In some examples, the path of the object comprises a hierarchy (e.g., directory hierarchy). In some examples, the path of the object does not comprise a hierarchy.

[071] The systems and methods for storing digital information can comprise one or more hashes. In some cases, the one or more hashes are determined using a hashing module. In some cases, the hashing module is executed on the one or more processing units, such as those described herein. In some cases, the hashing module comprises instructions for determining the one or more hashes (e.g., a hash function). In some cases, the instructions (e.g., a hash function) are stored on a memory, such as those described herein. In some cases, information comprising an object, a part of an object, or a pool item is stored using a hash. In some cases, a first one or more hashes of each of a one or more objects is determined and/or a second one or more hashes of each of a one or more pool items is determined. In some instances, a hash of a pool item is appended to the data payload. In some instances, a hash of an object is appended to the end pool descriptor.

[072] A hash may be determined a hash function (**FIG. 4**). A hash function generally comprises a function that turns an input of arbitrary length into an output with a fixed length (e.g., 224, 256, 384, 512 bits or characters). In some cases, the hash function comprises a cryptographic hash function. In some cases, the hash function comprises MD-5, SHA-1, SHA-2, SHA-3, RIPEMD-160, Whirlpool, BLAKE, BLAKE2, BLAKE3, or a variation thereof. In some instances, the hash function comprises SHA-2. In some examples, SHA-2 comprises SHA-224, SHA-256, SHA-384, SHA-512, SHA-512/224, or SHA-512/256. The output of a hash function can be deterministic and infeasible to reverse-engineer. Further, generating an output of fixed length can increase security, since any party involved in decrypting a hash would not be able to tell the length of the input. In some examples, a hash is generated upon inputting an identification code, encryption key, password, or any variation thereof. In some examples, the hash allows verification of the content (e.g., item of information or digital information stored in a pool) during decoding.

[073] In some cases, the input **405** comprises an object. In some examples, a hash function **410** is used to determine a hashed output (or hash) **415**. In some cases, the input **420** comprises an object. In some examples, a hash function **425** is used to determine a hashed output (or hash) **430**. In some examples, the hash function **410** and hash function **425** are the same hash function. In some examples, the hash function

410 and hash function **425** are both SHA-256. In some examples, the hash function **410** and hash function **425** are different hash functions. In some examples, the output **415** and the output **430** are the same length. In some examples, the output **415** and the output **430** are both 256 bits. In some examples, the output **415** and the output **430** are different lengths.

[074] A hash function can comprise one or more steps to generate a hash. In some cases, the one or more steps in a hash function comprises padding bits. In some instances, extra bits are added to the digital information (or the message) being hashed. In some examples, extra bits are added to the message such that the length of the digital message is a modulus value less than a total number of bits. In some examples, the modulus value is 64 bits. In some examples, the number of bits is 512 bits and the length of the digital information is 448 bits (e.g., for SHA-256). In some examples, the first extra bit comprises a binary digit of 1. In some examples, the subsequently added extra bits comprise a binary digit of 0s.

[075] In some cases, the one or more steps in a hash function comprises padding a length. In some instances, padding the length comprises adding a modulus value to the digital information (e.g., also referred to as a bi-endian (BE) integer). The modulus value or the BE integer generally represents the length of the original input comprising the original digital information in binary. In some examples, the modulus value is 64 bits. In some examples, 64 bits are added to the digital message of 448 bits, and the total number of bits is 512 bits (e.g., for SHA-256). In some instances, the modulus value is calculated by applying a modulus to the original digital information. As an example, if the original digital information is "hello world" in binary, the length of the original input is 88 bits, which is "1011000" in binary. As such, 0s followed by "1011000" are added to the end of the 448 bits of digital information such that the total number of bits is 512.

[076] In some cases, the one or more steps in the hash function comprises initializing one or more hash values or buffers. In some instances, 8 hash values or buffers are initialized. In some instances, the initialized hash values are hard-coded (e.g., constants). In some instances, the initialized hash values represent a first 32 bits of fractional part of the square roots of the first 8 primes (e.g., 2, 3, 5, 7, 11, 13, 17, 19). In some cases, the one or more steps in the hash function further comprises initializing round constants (or keys). In some instances, 64 round constants are initialized. In some examples, each of the 64 round constants represent the first 32 bits of the fractional parts of the cube roots of the first 64 primes (e.g., 2-311). In some instances, the 64 different round constants are stored in an array.

[077] In some cases, the one or more steps in the hash function comprises compression. In some instances, each block of information (e.g., every 512 bits) undergoes compression. During compression, each block of information undergoes a fixed number of rounds. In some instances, the number of rounds is 64. In some instances, compression is performed by a one-way compression function. In some instances, the one-way compression function is single block-length compression function. In some examples the compression function is a Davies-Meyer, Matyas-Meyer-Oseas, or Miyaguchi-Preneel compression function. In some instances, the one-way compression function is double block-length

compression function. In some examples the compression function is a MDC-2/Meyer–Schilling, MDC-4, or Hirose compression function. In some instances, the output from the compression function is less than the block of information. In some examples, the output has a length of 256 bits.

[078] In some cases, one or more of the hashes (e.g., hashes of pool item(s), hashes of object(s)) are calculated during storage of information. In some cases, all of the hashes (e.g., hashes of pool item(s), hashes of object(s)) are calculated during storage of information. In some examples, this allows stable low memory usage regardless of the size of the objects. In some cases, the first one or more hashes of each of the one or more objects require less memory than the one or more objects. In some cases, the second one or more hashes of each of the one or more pool items require less memory than one or more pool items. In some cases, the source data (e.g., item of information) is read only once. In some cases, each of the pools are written once without seeks. In some examples, this minimizes data transfers and latency.

[079] In some cases, the hashes described herein can serve one or more purposes. The one or more purposes can comprise, by way of non-limiting example, one or more of: verifying the integrity of one or more items of information (e.g., an object), signature generation and verification (e.g., for digital signatures), password verification, proof-of-work, or identifier for item of information.

[080] In some cases, an encryption and/or compression can further be added. In some examples, the encryption and/or compression is implemented with streaming application programmable interface (API). In some examples, this avoids the need to store intermediate results. In some cases, the digital information to be stored is already compressed, for example, to reduce data transfer costs. In some cases, the digital information to be stored is already encrypted, for example, for security reasons.

[081] The one or more steps in the instructions stored on the memory can further comprise creating a plurality of index pools. In some cases, the plurality of index pools contain only indices. In some cases, the index pools are used when retrieving the objects stored in the plurality of pools encoded in a plurality of polynucleotides. In some instances, index pools are sequenced and temporarily stored in digital storage systems (e.g. flash drives) to search for objects. In some examples, once a pool is identified, the plurality of polynucleotides encoding the pool is sequenced.

[082] In some cases, the one or more index pools comprise an index pool descriptor and/or a list of object indexing. In some instances, the index pool descriptor comprises a version, a pool ID, a size of a pool, a timestamp, or a combination thereof. In some examples, the pool ID comprises a unique ID of the pool. In some examples, the unique ID comprises a universal unique identifier (UUID). In some examples, the unique ID comprises a content ID. In some examples, the content ID comprises a digital fingerprinting system, which can be used to identify and/or manage copyright or ownership of a content. In some examples, the size of each of the plurality of index pools is about 1GB to about 1 TB. In some instances, the list of an object indexing comprises a path of an object, a hash of an object, a list of object fragments, a list of object metadata, or any combination thereof. In some examples, the path of the object

comprises a unique path. In some examples, the path of the object comprises a hierarchy (e.g., directory hierarchy). In some examples, the path of the object does not comprise a hierarchy. In some examples, the hash of the object is a hash as previously described herein (e.g., SHA-256). In some examples, the list of object fragments comprises a pool ID of a pool containing a fragment, a range of a fragment, or a combination thereof. In some examples, the list of object metadata comprises the type of metadata, the metadata payload, or an combination thereof. In some examples, the type of metadata comprises, a list of keywords attached to an object, a thumbnail picture, a text summary, an ID range for a sorted key-value database, a timestamp, a version, or any combination thereof. In some examples, the metadata is customizable. In some examples, the metadata is used to search for an object in the plurality of pools.

[083] In some cases, an index pool can store information of about 1 to about 1 million pools. In some cases, an index pool can store information of about 1 pool to about 10 pools, about 1 pool to about 100 pools, about 1 pool to about 1,000 pools, about 1 pool to about 5,000 pools, about 1 pool to about 10,000 pools, about 1 pool to about 50,000 pools, about 1 pool to about 100,000 pools, about 1 pool to about 500,000 pools, about 1 pool to about 1 million pools, about 10 pools to about 100 pools, about 10 pools to about 1,000 pools, about 10 pools to about 5,000 pools, about 10 pools to about 10,000 pools, about 10 pools to about 50,000 pools, about 10 pools to about 100,000 pools, about 10 pools to about 500,000 pools, about 10 pools to about 1 million pools, about 100 pools to about 1,000 pools, about 100 pools to about 5,000 pools, about 100 pools to about 10,000 pools, about 100 pools to about 50,000 pools, about 100 pools to about 100,000 pools, about 100 pools to about 500,000 pools, about 100 pools to about 1 million pools, about 1,000 pools to about 5,000 pools, about 1,000 pools to about 10,000 pools, about 1,000 pools to about 50,000 pools, about 1,000 pools to about 100,000 pools, about 1,000 pools to about 500,000 pools, about 1,000 pools to about 1 million pools, about 5,000 pools to about 10,000 pools, about 5,000 pools to about 50,000 pools, about 5,000 pools to about 100,000 pools, about 5,000 pools to about 500,000 pools, about 5,000 pools to about 1 million pools, about 10,000 pools to about 50,000 pools, about 10,000 pools to about 100,000 pools, about 10,000 pools to about 500,000 pools, about 10,000 pools to about 1 million pools, about 50,000 pools to about 100,000 pools, about 50,000 pools to about 500,000 pools, about 50,000 pools to about 1 million pools, about 100,000 pools to about 500,000 pools, about 100,000 pools to about 1 million pools, or about 500,000 pools to about 1 million pools. In some cases, an index pool can store information of about 1 pool, about 10 pools, about 100 pools, about 1,000 pools, about 5,000 pools, about 10,000 pools, about 50,000 pools, about 100,000 pools, about 500,000 pools, or about 1 million pools. In some cases, an index pool can store information of at least about 1 pool, about 10 pools, about 100 pools, about 1,000 pools, about 5,000 pools, about 10,000 pools, about 50,000 pools, about 100,000 pools, or about 500,000 pools. In some cases, an index pool can store information of at most about 10 pools, about 100 pools, about 1,000 pools, about 5,000 pools, about 10,000 pools, about 50,000 pools, about 100,000 pools, about 500,000 pools, or about 1 million pools.

[084] In some cases, each of the one or more index pools is about 1 GB to about 1 TB. In some cases, each of the plurality of pools is about 1 GB to about 1 TB. In some cases, each of the one or more index

pools is about 1 GB to about 10 GB, about 1 GB to about 50 GB, about 1 GB to about 100 GB, about 1 GB to about 500 GB, about 1 GB to about 1 TB, about 10 GB to about 50 GB, about 10 GB to about 100 GB, about 10 GB to about 500 GB, about 10 GB to about 1 TB, about 50 GB to about 100 GB, about 50 GB to about 500 GB, about 50 GB to about 1 TB, about 100 GB to about 500 GB, about 100 GB to about 1 TB, or about 500 GB to about 1 TB. In some cases, each of the one or more index pools is about 1 GB, about 10 GB, about 50 GB, about 100 GB, about 500 GB, or about 1 TB. In some cases, each of the one or more index pools is at least about 1 GB, about 10 GB, about 50 GB, about 100 GB, or about 500 GB. In some cases, each of the one or more index pools is at most about 10 GB, about 50 GB, about 100 GB, about 500 GB, or about 1 TB.

[085] Encoding Scheme

[086] An encoding scheme can be applied to each of the plurality of pools and/or index pools. In some cases, the encoding scheme encodes the digital information in the plurality of pools as a plurality of polynucleotides. In some cases, the encoding scheme encodes the digital information in the index pools as a plurality of polynucleotides. In some instances, the encoding scheme comprises codecs for encoding binary data as nucleic acid sequences (e.g., inner codec). In some instances, the encoding scheme comprises an error correction code (ECC) (e.g., outer codec). In some cases, the encoding scheme (e.g., inner codec or low-level codec) is also designed and implemented to allow streaming read and write API access. In some cases, the encoding scheme (e.g., inner codec or low-level codec) is also designed and implemented to match the streaming of the systems and methods for digital storage (e.g., outer codec or high-level codec) described herein.

[087] The encoding scheme can generally comprise one or more operations. The one or more operations can comprise one or more operation to manipulate or transform data (e.g., digital information). The one or more operations can comprise by way of non-limiting example, splitting, shuffling, concatenating, transposing, translating, duplicating, labeling (e.g., using an index) data or a part of the data, or any combination thereof.

[088] As an example, method of encoding digital or binary data in a plurality of nucleotide sequences can comprise splitting the binary data into a plurality of frames. In some instances, the plurality of frames comprise about 100 to about 10,000 frames. In some instances, the plurality of frames comprise about 100 frames to about 250 frames, about 100 frames to about 500 frames, about 100 frames to about 750 frames, about 100 frames to about 1,000 frames, about 100 frames to about 2,500 frames, about 100 frames to about 5,000 frames, about 100 frames to about 7,500 frames, about 100 frames to about 10,000 frames, about 250 frames to about 500 frames, about 250 frames to about 750 frames, about 250 frames to about 1,000 frames, about 250 frames to about 2,500 frames, about 250 frames to about 5,000 frames, about 250 frames to about 7,500 frames, about 250 frames to about 10,000 frames, about 500 frames to about 750 frames, about 500 frames to about 1,000 frames, about 500 frames to about 2,500 frames, about 500 frames to about 5,000 frames, about 500 frames to about 7,500 frames, about 500 frames to

about 10,000 frames, about 750 frames to about 1,000 frames, about 750 frames to about 2,500 frames, about 750 frames to about 5,000 frames, about 750 frames to about 7,500 frames, about 750 frames to about 10,000 frames, about 1,000 frames to about 2,500 frames, about 1,000 frames to about 5,000 frames, about 1,000 frames to about 7,500 frames, about 1,000 frames to about 10,000 frames, about 2,500 frames to about 5,000 frames, about 2,500 frames to about 7,500 frames, about 2,500 frames to about 10,000 frames, about 5,000 frames to about 7,500 frames, about 5,000 frames to about 10,000 frames, or about 7,500 frames to about 10,000 frames. In some instances, the plurality of frames comprise about 100 frames, about 250 frames, about 500 frames, about 750 frames, about 1,000 frames, about 2,500 frames, about 5,000 frames, about 7,500 frames, or about 10,000 frames. In some instances, the plurality of frames comprise at least about 100 frames, about 250 frames, about 500 frames, about 750 frames, about 1,000 frames, about 2,500 frames, about 5,000 frames, or about 7,500 frames. In some instances, the plurality of frames comprise at most about 250 frames, about 500 frames, about 750 frames, about 1,000 frames, about 2,500 frames, about 5,000 frames, about 7,500 frames, or about 10,000 frames. In some cases, the frames each comprise the same amount of data. In alternative cases, the frames each comprise a different amount of data. In some instances, each frame is assigned a frame index. In some examples, the frame index increases for each frame index (e.g., 0, 1, 2, 3, 4, 5, ..., etc.). In some examples, the frame index monotonically increases for each frame index.

[089] Methods for encoding digital or binary data comprise an outer codec. In some instances, methods for encoding digital or binary data in a plurality of nucleotide sequences comprise an outer codec. In some instances, an outer codec is applied to the binary data. In some instances, an outer codec is applied to the binary data once the binary data is split into a plurality of frames. In such instances, outer codec is applied to each of the plurality of frames. An exemplary diagram of splitting a data stream into frames and applying an outer codec is exemplary illustrated in **FIG. 5**.

[090] In some instances, the outer codec comprises an error correction code or scheme, such as a Reed-Solomon (RS) code. This outer codec is used for spreading the digital or binary data to be stored over many oligonucleotides. In some instances, spreading the data builds redundancy to correct for erasures (e.g., lost oligos). In some further embodiments, spreading the data also builds redundancy to correct errors from an inner codec.

[091] In some instances, the error correction scheme comprises Reed-Solomon (RS) code. In such instances, a RS encoder is used to encode the binary data or plurality of frames comprising binary data. Generally the RS codes operates on a block of data treated as a set of finite-field elements. In some instances, the RS code comprises mapping data, e.g., $x = (x_1, \dots, x_k) \in F^k$, to a polynomial p_x , where $p_x(a) = \sum_{i=1}^k x_i a^{i-1}$. The encoded data $C(x)$ is obtained by evaluating p_x at various different n points $a_1, \dots, a_n$ in the field F (e.g., $C(x) = (p_x(a_1), \dots, p_x(a_n))$).

[092] In some further embodiments, the RS code comprises an encoding scheme in which each codeword contains the message as a prefix, and error-correcting symbols are appended as a suffix. In

some instances, the RS code is specified as RS(n, k) with m -bit symbols. In such instances, the encoder takes k data symbols of m -bits each, and adds parity symbols (error correcting symbols or check symbols) to make an n symbol codeword. Here, there is $n - k$ parity symbols (or check symbols, t) of m bits each. In some cases, the RS decoder corrects up to t symbols that contain errors in a codeword, where $2t = n - k$. The codeword $C(x)$ comprises the parity check information $CK(x)$ which is systematically appended to the message information $M(x)$. The codeword $C(x)$ can be calculated as: $C(x) = x^{n-k} M(x) + CK(x) = x^{n-k} M(x) + x^{n-k} M(x) \bmod g(x)$. Here, k refers to the message length (e.g., symbols), t refers to the number of errors to be corrected, n refers to the block length (e.g., message length n plus the correction length t), and m refers to the symbol width, where given the symbol size, m , the maximum codeword length n for RS code is $n = 2^m - 1$. Further, x^{n-k} refers to the displacement shift in the message, and $g(x)$ refers to the generator polynomial, which is defined as the polynomial whose roots are sequential powers of the Galois field (GF) primitive α (e.g., $g(x) = (x - \alpha^i)(x - \alpha^{i+1}) \dots (x - \alpha^{i+n-k-1}) = g_0 + g_1x + \dots + g_{n-k-1}x^{n-k-1} + x^{n-k}$).

[093] For example, in RS(255, 223) with 8-bit symbols, the block length n is 255 codeword bytes, the message length k is 223 bytes, and the parity $2t$ is 32 bytes. In such an example, the RS decoder corrects up to 16 symbol errors in the codeword, meaning errors up to 16 bytes can be corrected by the decoder. The RS code can also be denoted based on the Galois Field as GF(2^m). For example, in an RS GF(2^{12}) encoding scheme, as shown in FIG. 5, n is 4095 (e.g., $n = 2^{12} - 1 = 4096 - 1 = 4095$). If k is, for example 2499, then $2t = 4095 - 2499 = 1596$ and t is thus be 798.

[094] In some instances, the error correction scheme comprises a linear error correction code (or linear block code), such as a low-density parity-check (LDPC) code. In some cases, the error correction scheme comprises a linear block error-correcting code, such as polar code. In some further embodiments, the error correction scheme comprises a high-performance forward error correction (FEC), such as a Turbo-code. In some instances, the error correction scheme comprises an RS code, an LDPC code, a Turbo-code, a polar code, or any combination thereof (e.g., RS-based LDPC codes).

[095] In some instances, the error correction scheme comprises low density parity check (LDPC) code. In such instances, the LDPC code is used to encode the binary data or plurality of frames comprising binary data. Generally the structure of a LDPC code is defined by a parity check matrix containing 0s at most entries and 1s elsewhere. For instance, an (N, K) LDPC code for K information bits is a linear block code with a block size of N , defined by a sparse $(N-K) \times N$ parity check matrix in which all elements other than 1s are 0s. The number of 1s in a row or a column is referred to as the degree of the row or the column. In some instances, a codeword of length N is represented as a vector C and for information bits of length K , an (N, K) code with $2K$ codewords is used. In some instances, the (N, K) LDPC code is defined by an $(N-K) \times N$ parity check matrix H , satisfying the condition: $HC^T = 0$.

[096] In some instances, the LDPC code is regular when each row and each column of the parity check matrix has a constant degree and irregular otherwise. In some instances, an irregular LDPC code

outperforms a regular LDPC code. In some instances, due to different degrees among rows and among columns, the irregular LDPC code promises improved performance only if the row degrees and the column degrees are appropriately adjusted.

[097] In some instances, the error correction scheme comprises a polar code. In some instances, a polar code can achieve Shannon capacity by theoretical proof. In some instances, a polar code comprises low encoding and decoding complexity. A polar code generally comprises a generator matrix G_N , and information can be encoded according to $x_I^N = u_I^N G^N$, where x_I^N is an output bit after encoding, u_I^N is an input bit before encoding, and the generator matrix is defined as $G^N = B_N F^{\otimes n}$. The code length N is defined as $N=2n$, where $n \geq 0$. B_N comprises a transposed matrix such as, for example, a bit reversal matrix. $F^{\otimes n}$ comprises a Kronecker power of F , which is defined as $F^{\otimes n} = F^{\otimes} F^{\otimes n-1}$, where $F = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$.

[098] In some instances, the Polar code is represented as (N, K, A, u_A^c) with a cosec code, and the encoding process is defined as $x_I^N = u_A G_N(A) \otimes u_A^c G_N(A^c)$, where A is an information bit index set, $G_N(A)$ is a submatrix obtained from a row, which corresponds to the index in the set A , in G_N . Further, $G_N(A^c)$ is a submatrix obtained from a row, which corresponds to the index in the set A^c , in G_N , and u_A^c is frozen bits the number of which is $(N-K)$, with N being the code length and K being the length of information bits. In some instances, the frozen bit is set to 0, and the above encoding process is described as $x_I^N = u_A G_N(A)$.

[099] In some instances, the error correction scheme comprises a turbo code. A turbo code generally comprises the parallel concatenation of two or more component codes applied to different interleaved versions of the same information sequence. Generally, recursive systematic convolutional (RSC) codes are used as the component codes. The structure of a turbo code, for example, comprises two RSC encoders (e.g., $M=2$), concatenated in parallel, and the code rate R is $R=1/3$, since $R=1/(M+1)$ (approximately). The input to the first RSC encoder is the original information sequence. The original information sequence d is also applied to an interleaver to produce an interleaved version d' . The interleaved version d' of the information sequence is the input to the second RSC encoder. The outputs from the turbo encoder comprise systematic sequences of u and redundant parts $x_{(1)}$ (output from the first RSC encoder) and $x_{(2)}$ (output from the second encoder). Therefore, the output of the encoder comprises $u_1, x_{1(1)}, x_{1(2)}, u_2, x_{2(1)}, x_{2(2)}$, where u_k is the k^{th} systematic bit (i.e., data bit), $x_{k(1)}$ is the parity output from the first RSC encoder associated with the k^{th} systematic bit u_k ; and $x_{k(2)}$ is the parity output from the second RSC encoder associated with the k^{th} systematic bit u_k . The decoding procedure for the turbo codes generally comprises iterative decoding. The turbo code decoding procedure can comprise two component decoders (corresponding to two RSC encoders), an interleaver; and, a de-interleaver. In some instances, the two component decoders are soft-input and soft-output (SISO) decoders. In some instances, outputs of the two component decoders comprise likelihood information concerning the coded data sequence.

[0100] In some instances, the size of the binary data is increased once an outer codec (e.g., ECC) is

applied. In some instances, the frame sizes are increased once an ECC is applied to each of the frame comprising binary data. In some instances, the frames are divided into a plurality of lanes. In some instances, each lane comprises a lane index. In some cases, each frame comprises about 1000 to about 10,000 lanes. In some cases, each frame comprises about 5000 lanes. In some cases, each frame comprises about 1,000 lanes to about 2,500 lanes, about 1,000 lanes to about 5,000 lanes, about 1,000 lanes to about 7,500 lanes, about 1,000 lanes to about 10,000 lanes, about 2,500 lanes to about 5,000 lanes, about 2,500 lanes to about 7,500 lanes, about 2,500 lanes to about 10,000 lanes, about 5,000 lanes to about 7,500 lanes, about 5,000 lanes to about 10,000 lanes, or about 7,500 lanes to about 10,000 lanes. In some cases, each frame comprises about 1,000 lanes, about 2,500 lanes, about 5,000 lanes, about 7,500 lanes, or about 10,000 lanes. In some cases, each frame comprises at least about 1,000 lanes, about 2,500 lanes, about 5,000 lanes, or about 7,500 lanes. In some cases, each frame comprises at most about 2,500 lanes, about 5,000 lanes, about 7,500 lanes, or about 10,000 lanes. Each lane can further comprise about 100 to about 300 bits. In some cases, each lane comprises about 100 bits to about 150 bits, about 100 bits to about 200 bits, about 100 bits to about 250 bits, about 100 bits to about 300 bits, about 150 bits to about 200 bits, about 150 bits to about 250 bits, about 150 bits to about 300 bits, about 200 bits to about 250 bits, about 200 bits to about 300 bits, or about 250 bits to about 300 bits. In some cases, each lane comprises about 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300 bits. In some cases, each lane comprises at least about 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300 bits. In some cases, each lane comprises at most about 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300 bits.

[0101] In some instances, the methods for encoding digital or binary data in a plurality of nucleotide sequences comprise shuffling the binary data. In some instances, each lane is shuffled base at least in part on lane indices. In some instances, each lane is shuffled after applying an outer codec (e.g., ECC) to the binary data. In some cases, shuffling each lane allows resistance against errors that can occur during synthesis or sequencing, such as those affecting a whole oligonucleotide library. The errors can comprise an insertion, a deletion, a substitution, or a combination thereof. In some instances, the shuffling comprises a rotation scheme within each lane based partly on each lane index. For example, each bit in a lane may be shifted by each lane index (e.g., no shuffling in lane 0, 1 bit shift in lane 1, 2 bit shift in lane 2, etc.).

[0102] In some further instances, the shuffling comprises a pseudorandom process within each lane. In this pseudorandom shuffling process, a random seed are used to initialize a pseudorandom number generator. In some instances, a number generated by the pseudorandom number generator is determined by the random seed. Therefore, the same sequence of numbers are generated by the pseudorandom number generator using the same seed. As an example, using shuffling comprises a pseudorandom process, each bit in a lane is be shifted according to the numbers generated by the pseudorandom number generator.

[0103] In some further instances, the lane index is used as a seed to create a permutation of some or all the bits for that lane. In some instances, the permutation of the some or all the bits is created by sampling from a random number generator. In some instances, the permutation is stored in a pre-compiled form. In some instances, the use of a pseudo random generator allows for a smaller implementation source code.

[0104] In some instances, the frame index and the lane index are prepended. In some instances, the frame index and the lane index are prepended to each lane once each lane is shuffled. An exemplary diagram of shuffling the lanes and prepending the frame index and the lane index is shown in **FIG. 6**. In some cases, the frame index comprises about 12 bits to about 20 bits. In some cases, the frame index comprises about 12 bits to about 14 bits, about 12 bits to about 16 bits, about 12 bits to about 18 bits, about 12 bits to about 20 bits, about 14 bits to about 16 bits, about 14 bits to about 18 bits, about 14 bits to about 20 bits, about 16 bits to about 18 bits, about 16 bits to about 20 bits, or about 18 bits to about 20 bits. In some cases, the frame index comprises about 12 bits, about 14 bits, about 16 bits, about 18 bits, or about 20 bits. In some cases, the frame index comprises at least about 12 bits, about 14 bits, about 16 bits, or about 18 bits. In some cases, the frame index comprises at most about 14 bits, about 16 bits, about 18 bits, or about 20 bits. In some cases, the lane index comprises about 12 bits to about 16 bits. In some cases, the lane index comprises about 12 bits to about 14 bits, about 12 bits to about 16 bits, or about 14 bits to about 16 bits. In some cases, the lane index comprises about 12 bits, about 14 bits, or about 16 bits. In some cases, the lane index comprises at least about 12 bits, or about 14 bits. In some cases, the lane index comprises at most about 14 bits, or about 16 bits. As shown in **FIG. 6**, in some instances, the lane index is 12 bits and the frame index is 20 bits. In some cases, the lane index is the symbol width m from the RS code.

[0105] In some instances, the methods for encoding digital or binary data in a plurality of nucleotide sequences comprise an inner codec. In some instances, the inner codec is applied to the binary data. In some instances, the inner codec is applied to the binary data from the ECC. In some instances, the inner codec is applied to the lanes of the binary data. In some instances, the inner codec is applied to the lanes of the binary data once the lanes have been shuffled.

[0106] In some instances, the encoding scheme comprises an inner codec. In some instances, an inner codec is applied to each lane to encode the binary data as a nucleotide sequence. The inner codec is used to transform digital or binary data into nucleotide bases. In some instances, the inner codec is capable of correcting deletion, substitution, or insertion errors, or any combination thereof. In some further embodiments, the inner codec is used to validate oligos and discard any suspicious oligos to avoid contaminating the outer decoding. The inner codec further encodes the indices (frame index and lane index), which can allow for efficient clustering during decoding.

[0107] In some instances, the encoding scheme adds redundancy across the plurality of oligonucleotide sequences. In some instances, the redundancy is about 5 % to about 10 %. In some instances, the redundancy is about 5 % to about 6 %, about 5 % to about 7 %, about 5 % to about 8 %, about 5 % to

about 9 %, about 5 % to about 10 %, about 6 % to about 7 %, about 6 % to about 8 %, about 6 % to about 9 %, about 6 % to about 10 %, about 7 % to about 8 %, about 7 % to about 9 %, about 7 % to about 10 %, about 8 % to about 9 %, about 8 % to about 10 %, or about 9 % to about 10 %. In some instances, the redundancy is about 5 %, about 6 %, about 7 %, about 8 %, about 9 %, or about 10 %. In some instances, the redundancy is at least about 5 %, about 6 %, about 7 %, about 8 %, or about 9 %. In some instances, the redundancy is at most about 61 GB %, about 7 %, about 8 %, about 9 %, or about 10 %. In some cases, this redundancy allows a library of oligos to be decoded in the presence of errors in the individual oligos, such as insertions, deletions, substitutions, or any combination thereof.

[0108] An exemplary diagram of an encoding scheme is shown in FIG. 7. In this exemplary diagram, the encoding scheme in the inner codec combines two or more of: bits from each lane, a bit history, and a bit position. In some instances, a model (e.g., adaptive model) is used to partition known bits into a context, and each context is mapped to a bit history. In some cases, the bit history is represented by an 8-bit state. In some instances, the bit history is updated each time a context is encountered, for example, through the use of a lookup table. A bit position comprises the least significant bit (LSB) from a bit index of the bits to encode. For example, if 100 bits encode a 100-mer oligonucleotide, a “bit index” refers to an index from 0 to 99 in the bits to encode. The LSB comprises the bit position in a binary integer representing the binary 1s place of the integer. In some instances, the LSB index is any length. In some instances, the LSB index is represented by a 4-bit state.

[0109] In some instances, the inner codec comprises generating base candidates for bits of the binary data. Base candidates are generated for the binary data using a lookup table, a hash, or a combination thereof. In some instances, the hash is determined using methods previously described herein. In some instances, the binary data comprises two or more of: bits from each lane, bit history, and a bit position. In some instances, the bit rate for encoding is about 1 bit per base to about 2 bits per base. In some instances, the bit rate for encoding is about 1 bit per base to about 1.1 bits per base, about 1 bit per base to about 1.2 bits per base, about 1 bit per base to about 1.3 bits per base, about 1 bit per base to about 1.4 bits per base, about 1 bit per base to about 1.5 bits per base, about 1 bit per base to about 1.6 bits per base, about 1 bit per base to about 1.7 bits per base, about 1 bit per base to about 1.8 bits per base, about 1 bit per base to about 1.9 bits per base, about 1 bit per base to about 2 bits per base, about 1.1 bits per base to about 1.2 bits per base, about 1.1 bits per base to about 1.3 bits per base, about 1.1 bits per base to about 1.4 bits per base, about 1.1 bits per base to about 1.5 bits per base, about 1.1 bits per base to about 1.6 bits per base, about 1.1 bits per base to about 1.7 bits per base, about 1.1 bits per base to about 1.8 bits per base, about 1.1 bits per base to about 1.9 bits per base, about 1.1 bits per base to about 2 bits per base, about 1.2 bits per base to about 1.3 bits per base, about 1.2 bits per base to about 1.4 bits per base, about 1.2 bits per base to about 1.5 bits per base, about 1.2 bits per base to about 1.6 bits per base, about 1.2 bits per base to about 1.7 bits per base, about 1.2 bits per base to about 1.8 bits per base, about 1.2 bits per base to about 1.9 bits per base, about 1.2 bits per base to about 2 bits per base, about 1.3 bits per base to about 1.4 bits per base, about 1.3 bits per base to about 1.5 bits per base, about 1.3 bits per base to about

1.6 bits per base, about 1.3 bits per base to about 1.7 bits per base, about 1.3 bits per base to about 1.8 bits per base, about 1.3 bits per base to about 1.9 bits per base, about 1.3 bits per base to about 2 bits per base, about 1.4 bits per base to about 1.5 bits per base, about 1.4 bits per base to about 1.6 bits per base, about 1.4 bits per base to about 1.7 bits per base, about 1.4 bits per base to about 1.8 bits per base, about 1.4 bits per base to about 1.9 bits per base, about 1.4 bits per base to about 2 bits per base, about 1.5 bits per base to about 1.6 bits per base, about 1.5 bits per base to about 1.7 bits per base, about 1.5 bits per base to about 1.8 bits per base, about 1.5 bits per base to about 1.9 bits per base, about 1.5 bits per base to about 2 bits per base, about 1.6 bits per base to about 1.7 bits per base, about 1.6 bits per base to about 1.8 bits per base, about 1.6 bits per base to about 1.9 bits per base, about 1.6 bits per base to about 2 bits per base, about 1.7 bits per base to about 1.8 bits per base, about 1.7 bits per base to about 1.9 bits per base, about 1.7 bits per base to about 2 bits per base, about 1.8 bits per base to about 1.9 bits per base, about 1.8 bits per base to about 2 bits per base, or about 1.9 bits per base to about 2 bits per base. In some instances, the bit rate for encoding is about 1 bit per base, about 1.1 bits per base, about 1.2 bits per base, about 1.3 bits per base, about 1.4 bits per base, about 1.5 bits per base, about 1.6 bits per base, about 1.7 bits per base, about 1.8 bits per base, about 1.9 bits per base, or about 2 bits per base. In some instances, the bit rate for encoding is at least about 1 bit per base, about 1.1 bits per base, about 1.2 bits per base, about 1.3 bits per base, about 1.4 bits per base, about 1.5 bits per base, about 1.6 bits per base, about 1.7 bits per base, about 1.8 bits per base, or about 1.9 bits per base. In some instances, the bit rate for encoding is at most about 1.1 bits per base, about 1.2 bits per base, about 1.3 bits per base, about 1.4 bits per base, about 1.5 bits per base, about 1.6 bits per base, about 1.7 bits per base, about 1.8 bits per base, about 1.9 bits per base, or about 2 bits per base. In some instances, the lookup table is used to map bits to nucleotides (e.g., A = 00, T = 10, C = 01, G = 11). In some instances, a hash comprises a function that can be used to map data of an arbitrary size (e.g., arbitrary number bits) to a fixed size value (e.g., a hashed value). In some examples, the hashed value is mapped to nucleotide sequences.

[0110] In some instances, the inner codec comprises a base repetition check. In some instances, the base repetition check is performed once the base candidates are selected. In some instances, the base repetition check checks for repetitions in two or more sequential bases. In some instances, the base repetition check substitutes one base for another if there are repetition in two or more sequential bases. In some instances, the lookup table or the hash is updated based on bases that were updated during the base repetition check. Further, after the base repetition check, the bit history is updated. In some instances, the frame index and/or lane index are incremented. In some instances, this process is repeated until sequences of all of the plurality of nucleotide sequences are determined.

[0111] In some instances, the inner codec further comprises performing GC filtering prior to synthesizing the plurality of the nucleotide sequences. In some cases, the GC filtering removes about 1% to about 10% of lanes in the plurality of lanes. In some cases, the GC filtering removes about 5% to about 10% of lanes in the plurality of lanes. In some cases, the GC filtering removes no lanes in the plurality of lanes. In some cases, the GC filtering removes about 1 %, about 2 %, about 3 %, about 4 %, about 5 %, about 6 %, about 7 %, about 8 %, about 9 %, about 10 %, about 11 %, about 12 %, about 13 %, about 14 %, about 15 %, about 16 %, about 17 %, about 18 %, about 19 %, about 20 %, about 21 %, about 22 %, about 23 %, about 24 %, about 25 %, about 26 %, about 27 %, about 28 %, about 29 %, about 30 %, about 31 %, about 32 %, about 33 %, about 34 %, about 35 %, about 36 %, about 37 %, about 38 %, about 39 %, about 40 %, about 41 %, about 42 %, about 43 %, about 44 %, about 45 %, about 46 %, about 47 %, about 48 %, about 49 %, about 50 %, about 51 %, about 52 %, about 53 %, about 54 %, about 55 %, about 56 %, about 57 %, about 58 %, about 59 %, about 60 %, about 61 %, about 62 %, about 63 %, about 64 %, about 65 %, about 66 %, about 67 %, about 68 %, about 69 %, about 70 %, about 71 %, about 72 %, about 73 %, about 74 %, about 75 %, about 76 %, about 77 %, about 78 %, about 79 %, about 80 %, about 81 %, about 82 %, about 83 %, about 84 %, about 85 %, about 86 %, about 87 %, about 88 %, about 89 %, about 90 %, about 91 %, about 92 %, about 93 %, about 94 %, about 95 %, about 96 %, about 97 %, about 98 %, about 99 %, about 100 %.

about 6 %, about 7 %, about 8 %, about 9 %, or about 10 %. In some cases, the GC filtering removes at least about 1 %, about 2 %, about 3 %, about 4 %, about 5 %, about 6 %, about 7 %, about 8 %, or about 9 %. In some cases, the GC filtering removes at most about 2 %, about 3 %, about 4 %, about 5 %, about 6 %, about 7 %, about 8 %, about 9 %, or about 10 %. In some cases, the plurality of nucleotide sequences comprises about 40% to about 60% GC content. In some cases, the plurality of nucleotide sequences comprises about 40 % to about 45 %, about 40 % to about 50 %, about 40 % to about 55 %, about 40 % to about 60 %, about 45 % to about 50 %, about 45 % to about 55 %, about 45 % to about 60 %, about 50 % to about 55 %, about 50 % to about 60 %, or about 55 % to about 60 % GC content. In some cases, the plurality of nucleotide sequences comprises about 40 %, about 45 %, about 50 %, about 55 %, or about 60 % GC content. In some cases, the plurality of nucleotide sequences comprises at least about 40 %, about 45 %, about 50 %, or about 55 % GC content. In some cases, the plurality of nucleotide sequences comprises at most about 45 %, about 50 %, about 55 %, or about 60 % GC content. In some cases, at least 90% of the plurality of nucleotide sequences comprises about 40% to about 60 % GC content. In some cases, at least 90% of the plurality of nucleotide sequences comprises about 40 % to about 45 %, about 40 % to about 50 %, about 40 % to about 55 %, about 40 % to about 60 %, about 45 % to about 50 %, about 45 % to about 55 %, about 45 % to about 60 %, about 50 % to about 55 %, about 50 % to about 60 %, or about 55 % to about 60 % GC content. In some cases, at least 90% of the plurality of nucleotide sequences comprises about 40 %, about 45 %, about 50 %, about 55 %, or about 60 % GC content. In some cases, at least 90% of the plurality of nucleotide sequences comprises at least about 40 %, about 45 %, about 50 %, or about 55 % GC content. In some cases, at least 90% of the plurality of nucleotide sequences comprises at most about 45 %, about 50 %, about 55 %, or about 60 % GC content. The output from the inner codec comprises an final oligonucleotide library.

[0112] An exemplary diagram of an alternative encoding scheme is shown in **FIG. 8**. In some instances, the encoding scheme in the inner codec comprises starting with a default lookup table. The default lookup table is used to select a word to encode within each lane. In some instances, the word is an 8 bit word or a byte. The lookup table is applied to generate base candidates for each word or byte) within each lane. A next lookup table is selected based on the previously encoded word or byte. In some instances, the encoding scheme further comprises performing a base repetition check, GC filtering, or a combination thereof, as previously described herein. In some instances, this process is repeated until sequences of all of the plurality of nucleotide sequences may be determined. The output from the inner codec comprises a final oligonucleotide library.

[0113] In some cases, the length of each of the oligonucleotides (or polynucleotides) in a library is about 20 to about 500 bases. In some cases, the length of each of the oligonucleotides (or polynucleotides) in a library is about 20 bases to about 50 bases, about 20 bases to about 100 bases, about 20 bases to about 200 bases, about 20 bases to about 300 bases, about 20 bases to about 400 bases, about 20 bases to about 500 bases, about 50 bases to about 100 bases, about 50 bases to about 200 bases, about 50 bases to about 300 bases, about 50 bases to about 400 bases, about 50 bases to about 500 bases, about 100 bases to

about 200 bases, about 100 bases to about 300 bases, about 100 bases to about 400 bases, about 100 bases to about 500 bases, about 200 bases to about 300 bases, about 200 bases to about 400 bases, about 200 bases to about 500 bases, about 300 bases to about 400 bases, about 300 bases to about 500 bases, or about 400 bases to about 500 bases. In some cases, the length of each of the oligonucleotides (or polynucleotides) in a library is about 20 bases, about 50 bases, about 100 bases, about 200 bases, about 300 bases, about 400 bases, or about 500 bases. In some cases, the length of each of the oligonucleotides (or polynucleotides) in a library is at least about 20 bases, about 50 bases, about 100 bases, about 200 bases, about 300 bases, or about 400 bases. In some cases, the length of each of the oligonucleotides (or polynucleotides) in a library is at most about 50 bases, about 100 bases, about 200 bases, about 300 bases, about 400 bases, or about 500 bases.

[0114] De Novo Polynucleotide Synthesis

[0115] Provided herein are systems and methods for synthesis of libraries of polynucleotides on a substrate. In some instances, the library comprising a plurality of polynucleotides from the encoding scheme are synthesized. In some examples, the library comprising the plurality of polynucleotides from the encoding scheme encode a pool of the plurality of pools. In some examples, the library comprising the plurality of polynucleotides from the encoding scheme encode an index pool. In some instances, methods comprise use of electrochemical deprotection. In some instances, the substrate is a flexible substrate. In some instances, at least 10^{10} , 10^{11} , 10^{12} , 10^{13} , 10^{14} , or 10^{15} bases are synthesized in one day. In some instances, at least 10×10^8 , 10×10^9 , 10×10^{10} , 10×10^{11} , or 10×10^{12} polynucleotides are synthesized in one day. In some cases, each polynucleotide synthesized comprises at least 20, 50, 100, 200, 300, 400 or 500 nucleobases. In some cases, these bases are synthesized with a total average error rate of less than about 1 in 100; 200; 300; 400; 500; 1000; 2000; 5000; 10000; 15000; 20000 bases. In some instances, these error rates are for at least 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%, or more of the polynucleotides synthesized. In some instances, these at least 90%, 95%, 98%, 99%, 99.5%, or more of the polynucleotides synthesized do not differ from a predetermined sequence for which they encode. In some instances, the error rate for synthesized polynucleotides on a substrate using the methods and systems described herein is less than about 1 in 200, less than about 1 in 1,000, less than about 1 in 2,000, less than about 1 in 3,000, or less than about 1 in 5,000. Individual types of error rates include mismatches, deletions, insertions, and/or substitutions for the polynucleotides synthesized on the substrate. The term “error rate” refers to a comparison of the collective amount of synthesized polynucleotide to an aggregate of predetermined polynucleotide sequences. In some instances, synthesized polynucleotides disclosed herein comprise a tether of 12 to 25 bases. In some instances, the tether comprises 10, 11, 12, 13, 14, 15, 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, 45, 46, 47, 48, 49, 50 or more bases.

[0116] Described herein are methods, systems, devices, and compositions wherein chemical reactions used in polynucleotide synthesis are controlled using electrochemistry. Electrochemical reactions in some

instances are controlled by any source of energy, such as light, heat, radiation, or electricity. For example, electrodes are used to control chemical reactions as all or a portion of discrete loci on a surface.

Electrodes in some instances are charged by applying an electrical potential to the electrode to control one or more chemical steps in polynucleotide synthesis. In some instances, these electrodes are addressable. Any number of the chemical steps described herein is in some instances controlled with one or more electrodes. Electrochemical reactions may comprise oxidations, reductions, acid/base chemistry, or other reaction that is controlled by an electrode. In some instances, electrodes generate electrons or protons that are used as reagents for chemical transformations. Electrodes in some instances directly generate a reagent such as an acid. In some instances, an acid is a proton. Electrodes in some instances directly generate a reagent such as a base. Acids or bases are often used to cleave protecting groups, or influence the kinetics of various polynucleotide synthesis reactions, for example by adjusting the pH of a reaction solution. Electrochemically controlled polynucleotide synthesis reactions in some instances comprise redox-active metals or other redox-active organic materials. In some instances, metal or organic catalysts are employed with these electrochemical reactions. In some instances, acids are generated from oxidation of quinones.

[0117] Control of chemical reactions with is not limited to the electrochemical generation of reagents; chemical reactivity may be influenced indirectly through biophysical changes to substrates or reagents through electric fields (or gradients) which are generated by electrodes. In some instances, substrates include but are not limited to nucleic acids. In some instances, electrical fields which repel or attract specific reagents or substrates towards or away from an electrode or surface are generated. Such fields in some instances are generated by application of an electrical potential to one or more electrodes. For example, negatively charged nucleic acids are repelled from negatively charged electrode surfaces. Such repulsions or attractions of polynucleotides or other reagents caused by local electric fields in some instances provides for movement of polynucleotides or other reagents in or out of region of the synthesis device or structure. In some instances, electrodes generate electric fields which repel polynucleotides away from a synthesis surface, structure, or device. In some instances, electrodes generate electric fields which attract polynucleotides towards a synthesis surface, structure, or device. In some instances, protons are repelled from a positively charged surface to limit contact of protons with substrates or portions thereof. In some instances, repulsion or attractive forces are used to allow or block entry of reagents or substrates to specific areas of the synthesis surface. In some instances, nucleoside monomers are prevented from contacting a polynucleotide chain by application of an electric field in the vicinity of one or both components. Such arrangements allow gating of specific reagents, which may obviate the need for protecting groups when the concentration or rate of contact between reagents and/or substrates is controlled. In some instances, unprotected nucleoside monomers are used for polynucleotide synthesis. Alternatively, application of the field in the vicinity of one or both components promotes contact of nucleoside monomers with a polynucleotide chain. Additionally, application of electric fields to a substrate can alter the substrates reactivity or conformation. In an exemplary application, electric fields

generated by electrodes are used to prevent polynucleotides at adjacent loci from interacting. In some instances, the substrate is a polynucleotide, optionally attached to a surface. Application of an electric field in some instances alters the three-dimensional structure of a polynucleotide. Such alterations comprise folding or unfolding of various structures, such as helices, hairpins, loops, or other 3-dimensional nucleic acid structure. Such alterations are useful for manipulating nucleic acids inside of wells, channels, or other structures. In some instances, electric fields are applied to a nucleic acid substrate to prevent secondary structures. In some instances, electric fields obviate the need for linkers or attachment to a solid support during polynucleotide synthesis.

[0118] A suitable method for polynucleotide synthesis on a substrate of this disclosure is a phosphoramidite-based synthesis of DNA. In some cases, a reagent for the phosphoramidite-based synthesis comprises any one of or a combination of a nucleoside phosphoramidite, an oxidizer, an activator, or a deblocker or the solvent comprises acetonitrile. In some instances, the phosphoramidite-based synthesis method comprises the controlled addition of a phosphoramidite building block, i.e. nucleoside phosphoramidite, to a growing polynucleotide chain in a coupling step that forms a phosphite triester linkage between the phosphoramidite building block and a nucleoside bound to the substrate. In some instances, the nucleoside phosphoramidite is provided to the substrate activated. In some instances, the nucleoside phosphoramidite is provided to the substrate with an activator. In some instances, nucleoside phosphoramidites are provided to the substrate in a 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100-fold excess or more over the substrate-bound nucleosides. In some instances, the addition of nucleoside phosphoramidite is performed in an anhydrous environment, for example, in anhydrous acetonitrile. Following addition and linkage of a nucleoside phosphoramidite in the coupling step, the substrate is optionally washed. In some instances, the coupling step is repeated one or more additional times, optionally with a wash step between nucleoside phosphoramidite additions to the substrate. In some instances, a polynucleotide synthesis method used herein comprises 1, 2, 3 or more sequential coupling steps. Prior to coupling, in many cases, the nucleoside bound to the substrate is de-protected by removal of a protecting group, where the protecting group functions to prevent polymerization. Protecting groups may comprise any chemical group that prevents extension of the polynucleotide chain. In some instances, the protecting group is cleaved (or removed) in the presence of an acid. In some instances, the protecting group is cleaved in the presence of a base. In some instances, the protecting group is removed with electromagnetic radiation such as light, heat, or other energy source. In some instances, the protecting group is removed through an oxidation or reduction reaction (e.g., a . In some instances, a protecting group comprises a triarylmethyl group. In some instances, a protecting group comprises an aryl ether. In some instances, a protecting group comprises a disulfide. In some instances a protecting group comprises an acid-labile silane. In some instances, a protecting group comprises an acetal. In some instances, a protecting group comprises a ketal. In some instances, a protecting group comprises an enol ether. In some instances, a protecting group comprises a methoxybenzyl group. In some instances, a protecting group comprises an azide. In some instances, a

protecting group is 4,4'-dimethoxytrityl (DMT). In some instances, a protecting group is a tert-butyl carbonate. In some instances, a protecting group is a tert-butyl ester. In some instances, a protecting group comprises a base-labile group.

[0119] Following coupling, phosphoramidite polynucleotide synthesis methods optionally comprise a capping step. In a capping step, the growing polynucleotide is treated with a capping agent. A capping step generally serves to block unreacted substrate-bound 5'-OH groups after coupling from further chain elongation, preventing the formation of polynucleotides with internal base deletions. Further, phosphoramidites activated with 1H-tetrazole often react, to a small extent, with the O6 position of guanosine. Without being bound by theory, upon oxidation with I₂/water, this side product, possibly via O6-N7 migration, undergoes depurination. The apurinic sites can end up being cleaved in the course of the final deprotection of the polynucleotide thus reducing the yield of the full-length product. The O6 modifications may be removed by treatment with the capping reagent prior to oxidation with I₂/water. In some instances, inclusion of a capping step during polynucleotide synthesis decreases the error rate as compared to synthesis without capping. As an example, the capping step comprises treating the substrate-bound polynucleotide with a mixture of acetic anhydride and 1-methylimidazole. Following a capping step, the substrate is optionally washed.

[0120] Following addition of a nucleoside phosphoramidite, and optionally after capping and one or more wash steps, a substrate described herein comprises a bound growing nucleic acid that may be oxidized. The oxidation step comprises oxidizing the phosphite triester into a tetracoordinated phosphate triester, a protected precursor of the naturally occurring phosphate diester internucleoside linkage. In some instances, phosphite triesters are oxidized electrochemically. In some instances, oxidation of the growing polynucleotide is achieved by treatment with iodine and water, optionally in the presence of a weak base such as a pyridine, lutidine, or collidine. Oxidation is sometimes carried out under anhydrous conditions using tert-Butyl hydroperoxide or (1S)-(+)-(10-camphorsulfonyl)-oxaziridine (CSO). In some methods, a capping step is performed following oxidation. A second capping step allows for substrate drying, as residual water from oxidation that may persist can inhibit subsequent coupling. Following oxidation, the substrate and growing polynucleotide is optionally washed. In some instances, the step of oxidation is substituted with a sulfurization step to obtain polynucleotide phosphorothioates, wherein any capping steps can be performed after the sulfurization. Many reagents are capable of the efficient sulfur transfer, including, but not limited to, 3-(Dimethylaminomethylidene)amino-3H-1,2,4-dithiazole-3-thione, DDTT, 3H-1,2-benzodithiol-3-one 1,1-dioxide, also known as Beaucage reagent, and N,N,N',N'-Tetraethylthiuram disulfide (TETD).

[0121] For a subsequent cycle of nucleoside incorporation to occur through coupling, a protected 5' end (or 3' end, if synthesis is conducted in a 5' to 3' direction) of the substrate bound growing polynucleotide is removed so that the primary hydroxyl group can react with a next nucleoside phosphoramidite. In some instances, the protecting group is DMT and deblocking occurs with trichloroacetic acid in

dichloromethane. In some instances, the protecting group is DMT and deblocking occurs with electrochemically generated protons. Conducting detritylation for an extended time or with stronger than recommended solutions of acids may lead to increased depurination of solid support-bound polynucleotide and thus reduces the yield of the desired full-length product. Methods and compositions described herein provide for controlled deblocking conditions limiting undesired depurination reactions. In some instances, the substrate bound polynucleotide is washed after deblocking. In some cases, efficient washing after deblocking contributes to synthesized polynucleotides having a low error rate.

[0122] Methods for the synthesis of polynucleotides on a substrate described herein may involve an iterating sequence of the following steps: application of a protected monomer to a surface of a substrate feature to link with either the surface, a linker or with a previously deprotected monomer; deprotection of the applied monomer so that it can react with a subsequently applied protected monomer; and application of another protected monomer for linking. One or more intermediate steps include oxidation and/or sulfurization. In some instances, one or more wash steps precede or follow one or all of the steps.

[0123] Methods for the synthesis of polynucleotides on a substrate described herein may comprise an oxidation step. For example, methods involve an iterating sequence of the following steps: application of a protected monomer to a surface of a substrate feature to link with either the surface, a linker or with a previously deprotected monomer; deprotection of the applied monomer so that it can react with a subsequently applied protected monomer; application of another protected monomer for linking, and oxidation and/or sulfurization. In some instances, one or more wash steps precede or follow one or all of the steps.

[0124] Methods for the synthesis of polynucleotides on a substrate described herein may further comprise an iterating sequence of the following steps: application of a protected monomer to a surface of a substrate feature to link with either the surface, a linker or with a previously deprotected monomer; deprotection of the applied monomer so that it can react with a subsequently applied protected monomer; and oxidation and/or sulfurization. In some instances, one or more wash steps precede or follow one or all of the steps.

[0125] Methods for the synthesis of polynucleotides on a substrate described herein may further comprise an iterating sequence of the following steps: application of a protected monomer to a surface of a substrate feature to link with either the surface, a linker or with a previously deprotected monomer; and oxidation and/or sulfurization. In some instances, one or more wash steps precede or follow one or all of the steps.

[0126] Methods for the synthesis of polynucleotides on a substrate described herein may further comprise an iterating sequence of the following steps: application of a protected monomer to a surface of a substrate feature to link with either the surface, a linker or with a previously deprotected monomer; deprotection of the applied monomer so that it can react with a subsequently applied protected monomer; and oxidation and/or sulfurization. In some instances, one or more wash steps precede or follow one or

all of the steps.

[0127] In some instances, polynucleotides are synthesized with photolabile protecting groups, where the hydroxyl groups generated on the surface are blocked by photolabile-protecting groups. When the surface is exposed to UV light, such as through a photolithographic mask, a pattern of free hydroxyl groups on the surface may be generated. These hydroxyl groups can react with photoprotected nucleoside phosphoramidites, according to phosphoramidite chemistry. A second photolithographic mask can be applied and the surface can be exposed to UV light to generate second pattern of hydroxyl groups, followed by coupling with 5'-photoprotected nucleoside phosphoramidite. Likewise, patterns can be generated and oligomer chains can be extended. Without being bound by theory, the lability of a photocleavable group depends on the wavelength and polarity of a solvent employed and the rate of photocleavage may be affected by the duration of exposure and the intensity of light. This method can leverage a number of factors such as accuracy in alignment of the masks, efficiency of removal of photoprotecting groups, and the yields of the phosphoramidite coupling step. Further, unintended leakage of light into neighboring sites can be minimized. The density of synthesized oligomer per spot can be monitored by adjusting loading of the leader nucleoside on the surface of synthesis.

[0128] The surface of a substrate described herein that provides support for polynucleotide synthesis may be chemically modified to allow for the synthesized polynucleotide chain to be cleaved from the surface. In some instances, the polynucleotide chain is cleaved at the same time as the polynucleotide is deprotected. In some cases, the polynucleotide chain is cleaved after the polynucleotide is deprotected. In an exemplary scheme, a trialkoxysilyl amine such as $(\text{CH}_3\text{CH}_2\text{O})_3\text{Si}-(\text{CH}_2)_2-\text{NH}_2$ is reacted with surface SiOH groups of a substrate, followed by reaction with succinic anhydride with the amine to create an amide linkage and a free OH on which the nucleic acid chain growth is supported. Cleavage includes gas cleavage with ammonia or methylamine. In some instances cleavage includes linker cleavage with electrically generated reagents such as acids or bases. In some instances, once released from the surface, polynucleotides are assembled into larger nucleic acids that are sequenced and decoded to extract stored information.

[0129] The surfaces described herein can be reused after polynucleotide cleavage to support additional cycles of polynucleotide synthesis. For example, the linker can be reused without additional treatment/chemical modifications. In some instances, a linker is non-covalently bound to a substrate surface or a polynucleotide. In some embodiments, the linker remains attached to the polynucleotide after cleavage from the surface. Linkers in some embodiments comprise reversible covalent bonds such as esters, amides, ketals, beta substituted ketones, heterocycles, or other group that is capable of being reversibly cleaved. Such reversible cleavage reactions are in some instances controlled through the addition or removal of reagents, or by electrochemical processes controlled by electrodes. Optionally, chemical linkers or surface-bound chemical groups are regenerated after a number of cycles, to restore reactivity and remove unwanted side product formation on such linkers or surface-bound chemical

groups.

[0130] Alternatively, the polymer synthesis can be enzymatic DNA synthesis. In some cases, the enzymatic DNA synthesis uses water as a solvent and the reagent is an enzyme terminal deoxynucleotidyl transferase (TdT) or a deblocker. In some cases, enzymatic synthesis of DNA uses a template-independent DNA polymerase, terminal deoxynucleotidyl transferase (TdT), which is a protein that evolved to rapidly catalyze the linkage of naturally occurring dNTPs. TdT adds nucleotides indiscriminately so it is stopped from continuing unregulated synthesis by various techniques such as tethering the TdT, creating variant enzymes, and using nucleotides that include reversible terminators to prevent chain elongation. TdT activity is maximized at approximately 37° C. and performs enzymatic reactions in an aqueous environment.

[0131] Devices for Polynucleotide Storage

[0132] The synthesized libraries of polynucleotides can be stored in device. In some cases, the device comprises a polynucleotide data storage system. In some cases, the libraries encoding pools (e.g., a plurality of pools or index pools) are stored in compartments. In some instances, the compartments comprise, by way of non-limiting example, active surfaces (e.g., loci), tubes, cells, spots, or any other physical storage solutions. In some examples, the compartments comprise locations (e.g., spots) on a microfluidic chip, such as a digital microfluidic chip. In some examples, the compartments are marked with a label. In some examples, the label comprises a barcode, a name (e.g., customer name, sample type, etc.), a timestamp, a list of objects stored, or any combination thereof.

[0133] In some cases, the device for storing digital information in DNA comprises one or more compartments. In some instances, each of the one or more compartments comprises a library comprising a plurality of polynucleotides. In some examples, the library encodes a pool comprising digital information corresponding to one or more objects (e.g., a pool of the plurality of pools described herein). In some examples, the pool comprises a pool descriptor, one or more pool items, an end pool descriptor, such as those described herein. In some examples, the pool comprises about 1 GB to about 1 TB of digital information, as previously described herein.

[0134] In some instances, each of the one or more compartments comprises a medium for storing the plurality of polynucleotides. In some examples, the medium comprises a solid, a liquid, a gas, or any combination thereof. In some examples, the medium comprises a salt solution. In some examples, the molar ratio of salt to DNA may range from about 20:1 to about 2:1. In some examples, the molar ratio depends on the molecular weight of the salt used and on the relative amounts of salt and DNA combined. In some examples, the molar ratio is calculated between the cation of the salt and the negatively charged phosphate groups of the DNA. In some examples, the salt solution comprises a molar ratio of less than 20:1 salt cation to phosphate groups in the DNA. In some examples, the salt solution is dried to create a dried product. In some cases, the salt solution comprises, by way of non-limiting examples, calcium chloride, calcium nitrate, calcium carbonate, calcium phosphate, magnesium chloride, magnesium

sulfate, magnesium nitrate, magnesium carbonate, lanthanum chloride, lanthanum nitrate, lanthanum carbonate, lanthanum bromide, or a mixture thereof. In some instances, the salt solution comprises barium (II) chloride dihydrate, calcium chloride dihydrate, copper (II) chloride anhydrous, lanthanum trichloride, magnesium dichloride hexahydrate, sodium chloride, or strontium chloride hexahydrate. In some instances, the concentration of the salt solution is about 0.01 nM to about 0.1 nM.

[0135] In some instances, a medium for storing the plurality of polynucleotides comprises nanoparticles. In some instances, the nanoparticles comprise silica nanoparticles. In some instances, a subset of the plurality of polynucleotides are encapsulated in the nanoparticles. In some instances, the nanoparticles encapsulating polynucleotides are stored in a water-free or near-to water-free environment. In some instances, nanoparticles comprise a protective layer of silica (e.g., tetraethoxysilane). In some instances, the nanoparticles comprise a co-interacting compound with the polynucleotides (e.g., N-[3-(Trimethoxysilyl)propyl]-N,N,N-trimethylammonium chloride). In some instances, the nanoparticles encapsulating polynucleotides are stored on a digital microfluidic chip. In some instances, the digital microfluidic chip allows for programmability of fluid. In some instances, the programmability allows for automated storage and/or retrieval of polynucleotides. In some instances, each location on a digital microfluidic chip comprises about 100 GB, 500 GB, 1 TB, 2 TB, 10 TB, 20 TB, 30 TB, or 50 TB. In some instances, each location comprises about 50 μ g, 100 μ g, 150 μ g, 200 μ g, 250 μ g, 300 μ g, 350 μ g, 400 μ g, 450 μ g, 500 μ g, 600 μ g, 700 μ g, 800 μ g, 900 μ g, or 1000 μ g of nanoparticles.

[0136] In some cases, each of the one or more compartments are in communication. In some instances, each of the one or more compartments are in communication through the medium. In some cases, each of the one or more compartments are not in communication. In some instances, each of the one or more compartments are not in communication through the medium.

[0137] In some cases, the device further comprises one or more second compartments. In some instances, each of the one or more second compartments comprises a second library. In some examples, the second library encodes an index pool, such as those described herein. In some cases, the one or more second compartments comprise a medium as previously described herein. In some cases, the one or more second compartments comprise the same medium as the one or more compartments. In some cases, the one or more second compartments comprise different media as the one or more compartments. In some cases, each of the one or more second compartments are in communication with each other and/or the one or more compartments (e.g., through the medium). In some cases, each of the one or more second compartments are not in communication with each other and/or the one or more compartments.

[0138] In some cases, the device further comprises a solid support comprising a surface. A such, described herein are devices for solid support based nucleic acid synthesis and storage, wherein the solid support has varying dimensions. In some instances, a size of the solid support is between about 40 and 120 mm by between about 25 and 100 mm. In some instances, a size of the solid support is about 80 mm by about 50 mm. In some instances, a width of a solid support is at least or about 10 mm, 20 mm, 40 mm,

60 mm, 80 mm, 100 mm, 150 mm, 200 mm, 300 mm, 400 mm, 500 mm, or more than 500 mm. In some instances, a height of a solid support is at least or about 10 mm, 20 mm, 40 mm, 60 mm, 80 mm, 100 mm, 150 mm, 200 mm, 300 mm, 400 mm, 500 mm, or more than 500 mm. In some instances, the solid support has a planar surface area of at least or about 100 mm²; 200 mm²; 500 mm²; 1,000 mm²; 2,000 mm²; 4,500 mm²; 5,000 mm²; 10,000 mm²; 12,000 mm²; 15,000 mm²; 20,000 mm²; 30,000 mm²; 40,000 mm²; 50,000 mm² or more. In some instances, the thickness of the solid support is between about 50 mm and about 2000 mm, between about 50 mm and about 1000 mm, between about 100 mm and about 1000 mm, between about 200 mm and about 1000 mm, or between about 250 mm and about 1000 mm. Non-limiting examples thickness of the solid support include 275 mm, 375 mm, 525 mm, 625 mm, 675 mm, 725 mm, 775 mm and 925 mm. In some instances, the thickness of the solid support is at least or about 0.5 mm, 1.0 mm, 1.5 mm, 2.0 mm, 2.5 mm, 3.0 mm, 3.5 mm, 4.0 mm, or more than 4.0 mm.

[0139] Described herein are devices wherein two or more solid supports are assembled. In some instances, solid supports are interfaced together on a larger unit. Interfacing may comprise exchange of fluids, electrical signals, or other medium of exchange between solid supports. This unit is capable of interface with any number of servers, computers, or networked devices. For example, a plurality of solid support is integrated onto a rack unit, which is conveniently inserted or removed from a server rack. The rack unit may comprise any number of solid supports. In some instances the rack unit comprises at least 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000, 50,000, 100,000 or more than 100,000 solid supports. In some instances, two or more solid supports are not interfaced with each other. Nucleic acids (and the information stored in them) present on solid supports can be accessed from the rack unit. Access includes removal of polynucleotides from solid supports, direct analysis of polynucleotides on the solid support, or any other method which allows the information stored in the nucleic acids to be manipulated or identified. Information in some instances is accessed from a plurality of racks, a single rack, a single solid support in a rack, a portion of the solid support, or a single locus on a solid support. In various instances, access comprises interfacing nucleic acids with additional devices such as mass spectrometers, HPLC, sequencing instruments, PCR thermocyclers, or other device for manipulating nucleic acids. Access to nucleic acid information in some instances is achieved by cleavage of polynucleotides from all or a portion of a solid support. Cleavage in some instances comprises exposure to chemical reagents (ammonia or other reagent), electrical potential, radiation, heat, light, acoustics, or other form of energy capable of manipulating chemical bonds. In some instances, cleavage occurs by charging one or more electrodes in the vicinity of the polynucleotides. In some instances, electromagnetic radiation in the form of UV light is used for cleavage of polynucleotides. In some instances, a lamp is used for cleavage of polynucleotides, and a mask mediates exposure locations of the UV light to the surface. In some instances, a laser is used for cleavage of polynucleotides, and a shutter opened/closed state controls exposure of the UV light to the surface. In some instances, access to nucleic acid information (including removal/addition of racks, solid supports, reagents, nucleic acids, or other component) is completely automated.

[0140] Solid supports as described herein comprise an active area. In some instances, the active area comprises regions, cells, features, or loci for nucleic acid synthesis. In some instances, the active area comprises regions or loci for nucleic acid storage. In some examples, the regions or loci comprise the one or more compartments. In some examples, the regions or loci comprise the second one or more compartments. In some instances, the regions are addressable. In some examples, the regions are addressable through an electrode.

[0141] The active area comprises varying dimensions. For example, the dimension of the active area is between about 1 mm to about 50 mm by about 1 mm to about 50 mm. In some instances, the active area comprises a width of at least or about 0.5, 1, 1.5, 2, 2.5, 3, 5, 5, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, or more than 80 mm. In some instances, the active area comprises a height of at least or about 0.5, 1, 1.5, 2, 2.5, 3, 5, 5, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, or more than 80 mm.

[0142] Described herein are devices for solid support based nucleic acid synthesis and storage, wherein the solid support has a number of sites (e.g., spots) or positions for synthesis or storage. In some instances, the solid support comprises up to or about 10,000 by 10,000 positions in an area. In some instances, the solid support comprises between about 1000 and 20,000 by between about 1000 and 20,000 positions in an area. In some instances, the solid support comprises at least or about 10, 30, 50, 75, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10,000, 12,000, 14,000, 16,000, 18,000, 20,000 positions by least or about 10, 30, 50, 75, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10,000, 12,000, 14,000, 16,000, 18,000, 20,000 positions in an area. In some instances the area is up to 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, or 2.0 inches squared. In some instances, the solid support comprises loci having a pitch of at least or about 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5, 6, 7, 8, 9, 10, or more than 10 μm . In some instances, the solid support comprises loci having a pitch of about 5 μm . In some instances, the solid support comprises loci having a pitch of about 2 μm . In some instances, the solid support comprises loci having a pitch of about 1 μm . In some instances, the solid support comprises loci having a pitch of about 0.2 μm . In some instances, the solid support comprises loci having a pitch of about 0.2 μm to about 10 μm , about 0.2 to about 8 μm , about 0.5 to about 10 μm , about 1 μm to about 10 μm , about 2 μm to about 8 μm , about 3 μm to about 5 μm , about 1 μm to about 3 μm or about 0.5 μm to about 3 μm . In some instances, the solid support comprises loci having a pitch of about 0.1 μm to about 3 μm .

[0143] The solid support for nucleic acid synthesis or storage as described herein comprises a high capacity for storage of data. For example, the capacity of the solid support is at least or about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more than 1000 petabytes. In some instances, the capacity of the solid support is between about 1 to about 10 petabytes or between about 1 to about 100 petabytes. In some instances, the capacity of the solid support is about 100 petabytes. In some instances, the data is stored as arrays of packets as droplets. In some examples, the

arrays of packets are addressable packets. In some examples, the packets are addressable using an electrode. In some instances, the data is stored as arrays of packets as droplets on a spot. In some instances, the data is stored as arrays of packets as dry wells. In some instances, the arrays comprise at least or about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, or more than 200 gigabytes of data. In some instances, the arrays comprise at least or about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200, or more than 200 terabytes of data. In some instances, an item of information is stored in a background of data. For example, an item of information encodes for about 10 to about 100 megabytes of data and is stored in 1 petabyte of background data. In some instances, an item of information encodes for at least or about 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 300, 400, 500, or more than 500 megabytes of data and is stored in 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 300, 400, 500, or more than 500 petabytes of background data.

[0144] Provided herein are devices for solid support based nucleic acid synthesis and storage, wherein following synthesis, the polynucleotides are collected in packets as one or more droplets. In some instances, the polynucleotides are collected in packets as one or more droplets and stored. In some instances, a number of droplets is at least or about 1, 10, 20, 50, 100, 200, 300, 500, 1000, 2500, 5000, 75000, 10,000, 25,000, 50,000, 75,000, 100,000, 1 million, 5 million, 10 million, 25 million, 50 million, 75 million, 100 million, 250 million, 500 million, 750 million, or more than 750 million droplets. In some instances, a droplet volume comprises 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, or more than 100 μm (micrometer) in diameter. In some instances, a droplet volume comprises 1-100 μm , 10-90 μm , 20-80 μm , 30-70 μm , or 40-50 μm in diameter.

[0145] In some instances, the polynucleotides that are collected in the packets comprise a similar sequence. In some instances, the polynucleotides further comprise a non-identical sequence to be used as a tag or barcode. For example, the non-identical sequence is used to index the polynucleotides stored on the solid support and to later search for specific polynucleotides based on the non-identical sequence. Exemplary tag or barcode lengths include barcode sequences comprising, without limitation, about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25 or more bases in length. In some instances, the tag or barcode comprise at least or about 10, 50, 75, 100, 200, 300, 400, or more than 400 base pairs in length.

[0146] Provided herein are devices for solid support based nucleic acid synthesis and storage, wherein the polynucleotides are collected in packets comprising redundancy. For example, the packets comprise about 100 to about 1000 copies of each polynucleotide. In some instances, the packets comprise at least or about 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, or more than 2000 copies of each polynucleotide. In some instances, the packets comprise about 1000X to about 5000X synthesis redundancy. Synthesis redundancy in some instances is at least or about 500X, 1000X, 1500X, 2000X, 2500X, 3000X, 3500X, 4000X, 5000X, 6000X, 7000X, 8000X, or more than 8000X. The polynucleotides that are synthesized using solid support based methods as described herein comprise various lengths. In some instances, the polynucleotides are synthesized and further stored on the

solid support. In some instances, the polynucleotide length is in between about 100 to about 1000 bases. In some instances, the polynucleotides comprise at least or about 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, or more than 2000 bases in length.

[0147] Sequencing

[0148] Polynucleotides are extracted and/or amplified from surfaces where they are synthesized or stored. After extraction and/or amplification of polynucleotides from the surface of a structure, suitable sequencing technology may be employed to sequence the polynucleotides. In some cases, the DNA sequence is read on the substrate or within a feature of a structure. In some cases, the polynucleotides stored on the substrate are extracted is optionally assembled into longer nucleic acids and then sequenced.

[0149] Polynucleotides synthesized and stored on the structures described herein encode data that can be interpreted by reading the sequence of the synthesized polynucleotides and converting the sequence into binary code readable by a computer. In some cases the sequences require assembly, and the assembly step may need to be at the nucleic acid sequence stage or at the digital sequence stage.

[0150] Provided herein are detection systems comprising a device capable of sequencing stored polynucleotides, either directly on the structure and/or after removal from the main structure. In cases where the structure is a reel-to-reel tape of flexible material, the detection system comprises a device for holding and advancing the structure through a detection location and a detector disposed proximate the detection location for detecting a signal originated from a section of the tape when the section is at the detection location. In some instances, the signal is indicative of a presence of a polynucleotide. In some instances, the signal is indicative of a sequence of a polynucleotide (e.g., a fluorescent signal). In some instances, information encoded within polynucleotides on a continuous tape is read by a computer as the tape is conveyed continuously through a detector operably connected to the computer. In some instances, a detection system comprises a computer system comprising a polynucleotide sequencing device, a database for storage and retrieval of data relating to polynucleotide sequence, software for converting DNA code of a polynucleotide sequence to binary code, a computer for reading the binary code, or any combination thereof.

[0151] Provided herein are sequencing systems that can be integrated into the devices described herein. Various methods of sequencing are well known in the art, and comprise “base calling” wherein the identity of a base in the target polynucleotide is identified. In some instances, polynucleotides synthesized using the methods, devices, compositions, and systems described herein are sequenced after cleavage from the synthesis surface. In some instances, sequencing occurs during or simultaneously with polynucleotide synthesis, wherein base calling occurs immediately after or before extension of a nucleoside monomer into the growing polynucleotide chain. Methods for base calling include measurement of electrical currents/voltages generated by polymerase-catalyzed addition of bases to a template strand. In some instances, synthesis surfaces comprise enzymes, such as polymerases. In some

instances, such enzymes are tethered to electrodes or to the synthesis surface. In some instances, enzymes comprise terminal deoxynucleotidyl transferases, or variants thereof.

[0152] Systems and Methods for Digital Information Retrieval

[0153] Provided herein are methods and systems for retrieval of digital information. In some cases, the digital information comprises one or more objects as previously described herein. In some cases, each of the one or more objects is about 1 GB to about 1 TB as previously described herein. In some cases, the one or more objects comprises an item of information, such as, but not limited to, those described herein.

[0154] In some cases, the systems and methods decode nucleotide sequences (e.g., polynucleotides, oligonucleotides, plurality of polynucleotides, etc.). In some cases, a method for retrieving a digital information stored in a plurality of polynucleotides comprises one or more steps.

[0155] In some cases, retrieving a digital information stored in a plurality of polynucleotides comprises accessing an index pool. In some instances, accessing an index pool comprises fully or partially sequencing an library encoding an index pool. In some examples, the index pool is encoded in the library using the systems and methods described herein. In some examples, the polynucleotides in a library encoding an index pool are sequenced using the systems and methods described herein. In some instances, more than one index pool are accessed. In some instances, the polynucleotides in more than one library are sequenced. In some instances, the sequenced library is temporarily stored in a memory storage system (e.g. flash drives). In some instances, the sequenced library is converted to digital information to retrieve an index pool. In some instances, the index pool is temporarily stored in a memory storage system (e.g. flash drives). In some instances, the digital information in the index pool is used to search for one or more objects of interest. In some examples, the one or more objects of interest are stored in a library comprising a plurality of polynucleotides encoding the one or more objects. In some examples, the one of more objects of interest are searched using a metadata associated with the one or more object. In some instances, accessing an index pool determines a plurality of pools corresponding to one or more objects.

[0156] In some cases, the one or more objects of interest is retrieved from a compartment in a storage device. In some cases, retrieving a digital information stored in a plurality of polynucleotides comprises sequencing the plurality of polynucleotides corresponding to one or more objects in a plurality of pools. In some instances, the plurality of polynucleotides are in a library. In some instances, the library is in a compartment of a device, as previously described herein. In some instances, the plurality of polynucleotides in a library encoding a pool are sequenced using the systems and methods described herein. In some cases, the pool is encoded in the library using the systems and methods described herein. In some instances, the plurality of polynucleotides in more than one compartment is sequenced to retrieve the one or more objects.

[0157] In some cases, retrieving a digital information stored in a plurality of polynucleotides further

comprises applying a decoding scheme. In some instances, the decoding scheme decodes the digital information in the plurality of pools. In some instances, the decoding scheme is applied to the sequenced library comprising a plurality of polynucleotides. In some instances, a decoding scheme comprises an inner codec, an outer codec (e.g., ECC), or a combination thereof. In some instances, the decoding scheme decodes a plurality of nucleotide sequences to generate an output comprising digital information (e.g., an object). In some instances, the decoding scheme comprises undoing operations in the encoding scheme. In some examples, the operations comprise, splitting, shuffling, concatenating, transposing, translating, duplicating, labeling (e.g., using an index) data or a part of the data, or any combination thereof.

[0158] Provided herein are methods and systems for decoding. In some cases, the methods and systems decode nucleotide sequences (e.g., polynucleotides, oligonucleotides, plurality of polynucleotides, etc.). In some instances, the nucleotide sequences are encoded using the methods described herein. In some instances, the methods and systems comprise an inner codec, an outer codec, or a combination thereof. In some instances, methods for decoding the plurality of nucleotide sequences may comprises determining the plurality of nucleotide sequences. In some cases, determining the plurality of nucleotide sequences comprises sequencing the nucleotides. In some instances, the nucleotides are sequenced using the methods described herein.

[0159] After sequencing the plurality of nucleotides, the encoded binary data is decoded. In some instances, the plurality of nucleotides are decoded using the schematic illustrated, by way of non-limiting example, in **FIG. 9**. The output from sequencing comprises an unordered list of reads (e.g., nucleotide sequences), as shown in **FIG. 9**.

[0160] In some instances, the sequenced polynucleotides, such as an unordered list of reads, are clustered after sequencing. In some cases, clustering is performed prior to applying an inner codec. In some instances, the sequenced polynucleotides are clustered based on an index, such as the frame index, the lane index, or a combination thereof. In such instances, the sequenced polynucleotides are partially decoded to obtain the frame index, the lane index, or the combination thereof. In some instances, clustering is performed using a hash function, as previously described herein. In some instances, a hash function is used if the bases in the nucleotide sequences were determined using a hash in the encoding scheme, as previously described herein.

[0161] In some instances, the sequenced polynucleotides (e.g., reads) are aligned. In some instances, the sequenced polynucleotides are aligned after they have been clustered. In some cases, the sequenced polynucleotides are aligned prior to applying the inner codec. In some instances, aligning comprises analyzing consensus of the reads (e.g., nucleotide sequences) using an alignment algorithm. In some examples, the alignment algorithm comprises a pairwise alignment algorithm, a multi-sequence alignment algorithm, or a combination thereof.

[0162] In some instances, a pairwise alignment algorithm comprises initializing a position for each read.

Initializing comprises aligning a nucleotide sequence to a position 0. Consensus of a next one or more bases are analyzed between reads. In some instances, about 3 to about 10 reads are analyzed for consensus. In some instances, about 3 to about 4, about 3 to about 5, about 3 to about 6, about 3 to about 7, about 3 to about 8, about 3 to about 9, about 3 to about 10, about 4 to about 5, about 4 to about 6, about 4 to about 7, about 4 to about 8, about 4 to about 9, about 4 to about 10, about 5 to about 6, about 5 to about 7, about 5 to about 8, about 5 to about 9, about 5 to about 10, about 6 to about 7, about 6 to about 8, about 6 to about 9, about 6 to about 10, about 7 to about 8, about 7 to about 9, about 7 to about 10, about 8 to about 9, about 8 to about 10, or about 9 to about 10 reads are analyzed for consensus. In some instances, about 3, about 4, about 5, about 6, about 7, about 8, about 9, or about 10 reads are analyzed for consensus. In some instances, at least about 3, about 4, about 5, about 6, about 7, about 8, or about 9 reads are analyzed for consensus. In some instances, at most about 4, about 5, about 6, about 7, about 8, about 9, or about 10 reads are analyzed for consensus. In some instances, the next one or more bases comprise the next 2 to 10 bases. In some instances, the next one or more bases is about 2, 3, 4, 5, 6, 7, 8, 9, or 10 bases. In some instances, the next one or more bases is at least about 2, 3, 4, 5, 6, 7, 8, or 9 bases. In some instances, the next one or more bases is at most about 3, 4, 5, 6, 7, 8, 9, or 10 bases. In some instances, the next one or more bases is about 2, 3, 4, or 5 bases. The consensus is analyzed between the reads, and it is determined whether the next one or more bases are correct. If there is consensus between a base at a position, e.g., x , between all reads, then the subsequent base, e.g., $x+1$, may then be analyzed. If there is a inconsistencies in a base at a position, e.g., x , among the reads, then it is determined whether the read comprising the inconsistency has an error. In some instances, the error is an insertion, deletion, or substitution. The position is then incremented, e.g., $x+1$, given the decision (e.g., whether it is correct or has an error) for each read. In some instances, the steps are repeated until the end of a read is reached.

[0163] In some instances, decoding scheme comprise an inner codec. In some instances, the inner codec is applied to the plurality of nucleotide sequences. The inner codec is used to transform the nucleotide sequences into digital or binary data. In some instances, the inner codec is capable of correcting deletion, substitution, or insertion errors, or any combination thereof. In some further embodiments, the inner codec is used to validate oligos and discard any suspicious oligos to avoid contaminating the outer decoding. In some instances, the inner codec allows for efficient decoding using the indices (frame index and lane index).

[0164] An inner codec comprising a decoding scheme is applied to the plurality of nucleotide sequences. In some instances, the inner codec may transform each of the plurality of nucleotide sequences into lanes of binary data. In some instances, the inner codec is applied to a plurality of nucleotides that have been sequenced. In some instances, the inner codec is applied to the unordered reads. In some instances, the inner codec is applied to the reads or the plurality of polynucleotides once they have been clustered, as described herein. In some instances, the inner codec is applied to the reads or the plurality of nucleotides once they have been aligned, as described herein.

[0165] In some instances, the inner codec comprises a greedy algorithm. In some instances, the inner codec comprises a maximum likelihood (ML) algorithm. In some instances, the inner codec comprises a mixed greedy ML algorithm.

[0166] A inner codec comprising a greedy algorithm (e.g., greedy decoder) is exemplary illustrated in **FIG. 10**. As shown, a greedy algorithm takes into account transitions from only the most probably state as it decodes each bit position in a sequence. In some instances, each bit is guessed using the greedy algorithm one at a time. In some instances, more than one bit is guessed using the greedy algorithm at a given time. In some instances, the x-axis comprises the bit position and the y-axis comprises a state. In some instances, a state comprises one or more valid encoding states *S* that are analyzed at each bit position. In some instances, each state *S* is assigned a probability. In some instances, the state *S* is defined as the encoded bits from each lane, a bit history, and a bit position. In some instances, the state *S* is defined as the bit history and the bit word. The greedy algorithm repeatedly finds the highest probable state at each position until the highest probable end state is reached. In some instances, the decoded bits are backtracked by following the highest probable states at each bit position. In some instances, this results in the fully decoded bit. In some cases, the greedy decoder finds a locally optimal solution. In some instances, the locally optional solution is an approximate of a globally optimal solution. The greedy decoder provides a solution (or end state) in a reasonable amount of time compared to other inner codecs, such as those described herein.

[0167] In some instances, performance of the inner codec is improved by knowing where the oligonucleotide sequence ends. In some cases, the oligonucleotide lengths are determined during sequencing, for example, through pair-end sequencing. In some instances, a drift term is introduced to the greedy algorithm. The drift term comprises an integer associated with the total number of insertions and deletions. Each insertion is represented as a +1 value and each deletion is represented as a -1 value. For example, if there are no insertions and 2 deletions, the total drift is -2. In such an example, the greedy algorithm discards all end decoding states that do not match the length of oligo as being invalid. Therefore, the drift term allows the greedy algorithm to know which end decoding states are valid, and can further improve the performance. As such, in some instances, as shown in **FIG. 10** and **FIG. 11**, the inner codec further comprises a z-axis corresponding to the drift.

[0168] A inner codec comprising a ML algorithm is exemplary illustrated in **FIG. 11**. As shown, a ML algorithm takes into account transitions from all states as it decodes each bit position in a sequence. The states are defined as previously described herein. In some instances, each bit is guessed using the ML algorithm one at a time. In some instances, more than one bit is guessed using the ML algorithm at a given time. In some cases, the ML algorithm repeatedly finds all transition states at each position until end candidate states are determined. In some instances, the x-axis comprises the bit position and the y-axis comprises a state, as previously described herein. In some instances, a drift term, as previously described herein, is used to filter the end candidate states. In some instances, the ML algorithm provides

the globally optimal solution by tracking all state transitions. In some cases, the ML algorithm is computationally intensive compared to other decoding schemes, such as those described herein.

[0169] In some instances, an inner codec comprises a mixed greedy ML algorithm. A mixed greedy ML algorithm takes into account transitions from a plurality of states as it decodes each bit position in a sequence. In some instances, the plurality of states are about 100 to about 1000 states as it decodes each bit position in a sequence. In some instances, the plurality of states are about 100 to about 200, about 100 to about 300, about 100 to about 400, about 100 to about 500, about 100 to about 600, about 100 to about 700, about 100 to about 800, about 100 to about 900, about 100 to about 1,000, about 200 to about 300, about 200 to about 400, about 200 to about 500, about 200 to about 600, about 200 to about 700, about 200 to about 800, about 200 to about 900, about 200 to about 1,000, about 300 to about 400, about 300 to about 500, about 300 to about 600, about 300 to about 700, about 300 to about 800, about 300 to about 900, about 300 to about 1,000, about 400 to about 500, about 400 to about 600, about 400 to about 700, about 400 to about 800, about 400 to about 900, about 400 to about 1,000, about 500 to about 600, about 500 to about 700, about 500 to about 800, about 500 to about 900, about 500 to about 1,000, about 600 to about 700, about 600 to about 800, about 600 to about 900, about 600 to about 1,000, about 700 to about 800, about 700 to about 900, about 700 to about 1,000, about 800 to about 900, about 800 to about 1,000, or about 900 to about 1,000 states. In some instances, the plurality of states are about 100, about 200, about 300, about 400, about 500, about 600, about 700, about 800, about 900, or about 1,000 states. In some instances, the plurality of states are at least about 100, about 200, about 300, about 400, about 500, about 600, about 700, about 800, or about 900 states. In some instances, the plurality of states are at most about 200, about 300, about 400, about 500, about 600, about 700, about 800, about 900, or about 1,000 states. The states are defined as previously described herein. In some instances, each bit is guessed using the mixed greedy ML algorithm one at a time. In some instances, more than one bit is guessed using the mixed greedy ML algorithm at a given time. In some instances, the mixed greedy ML algorithm repeatedly finds about 100 to about 1000 transition states at each position until end candidate states are determined. In some instances, a drift term, as previously described herein, is used to filter the end candidate states. In some instances, the mixed greedy ML algorithm provides the globally optimal solution, while being less computationally expensive relative to other inner codecs, such as the ML algorithm described herein.

[0170] In some instances, the inner codec comprises a beam search decoder or a random sampling decoder (e.g., pure sampling decoder, a top-K sampling decoder, etc.). In some cases, a beam search decoder or a random sampling decoder provides a diversity of candidate states compared to a greedy decoder.

[0171] In some instances, the inner codec further comprises a checksum. In some instances, the checksum is used to verify data integrity, detect errors, or a combination thereof. In some instances, a checksum is generated using a checksum function or checksum algorithm (e.g., parity byte or parity work

(longitudinal parity check), sum complement, position dependent, fuzzy checksum, etc.). Examples of checksum functions or algorithms, include, but are not limited to, BSD checksum (Unix), SYSV checksum (Unix), sum4, sum8, sum16, sum32, fletcher-4, fletcher-8, fletcher-16, fletcher-32, Adler-32, xor8, Luhn algorithm, Verhoeff algorithm, or Damm algorithm. In some instances, instead of only taking the highest probable path, a few of the best probable paths are considered and tested against the checksum. In some instances, the checksum comprises a RS code (e.g., a small RS code). In such instances, the decoder gives a list of possibilities (e.g., “list decoding”) assuming the user can decide which one it actually is.

[0172] In some instances, decoding scheme further comprises arranging lanes into frames. In some instances, the decoded lanes from the inner codec are arranged into frames based on the lane index and the frame index. In some instances, one or more lanes are missing from a frame, as shown in **FIG. 9**. In some cases, the lanes are missing due to errors occurred during synthesis or sequencing of the nucleotides. In some cases, about 1% to about 10% of the lanes are missing from a frame. In some cases, about 1 % to about 2 %, about 1 % to about 4 %, about 1 % to about 6 %, about 1 % to about 8 %, about 1 % to about 10 %, about 2 % to about 4 %, about 2 % to about 6 %, about 2 % to about 8 %, about 2 % to about 10 %, about 4 % to about 6 %, about 4 % to about 8 %, about 4 % to about 10 %, about 6 % to about 8 %, about 6 % to about 10 %, or about 8 % to about 10 % of the lanes are missing from a frame. In some cases, about 1 %, about 2 %, about 4 %, about 6 %, about 8 %, or about 10 % of the lanes are missing from a frame. In some cases, at least about 1 %, about 2 %, about 4 %, about 6 %, or about 8 % of the lanes are missing from a frame. In some cases, at most about 2 %, about 4 %, about 6 %, about 8 %, or about 10 % of the lanes are missing from a frame.

[0173] In some instances, the inner codec comprises a “format”. In some cases, there is no a-priori information about the size of the data (e.g., binary data) during decoding. Therefore, in some instances, frame index 0 comprises the size of the data. In some instances, after arranging the lanes into frames and/or order the frames, frame 0 is decoded first. The data is then extracted from frame 0 to reject frames outside of the expected data size (e.g., from incorrectly decoded oligos).

[0174] In some instances, the inner codec comprises a hash (e.g., SHA-256). In some instances, the hash verifies that the data was correctly decoded. In some instances, by using a hash at the end (after the ECC), the encoding and decoding are performed as a stream. In some instances, this can limit memory use to only temporary buffers.

[0175] Methods for decoding a plurality of nucleotide sequences can comprise an outer codec (e.g., ECC). In some instances, the plurality of nucleotide sequences are decoded into digital or binary data. In some instances, an outer codec (e.g., ECC) is applied to the digital or binary data. In some examples, an ECC is applied to each of the frames. In some instances, the ECC is applied to the lanes from the inner codec. In some instances, the ECC is applied after the lanes from the inner codec are arranged into frames.

about 0.001 % to about 25 %, about 0.001 % to about 30 %, about 0.01 % to about 0.1 %, about 0.01 % to about 0.5 %, about 0.01 % to about 1 %, about 0.01 % to about 2 %, about 0.01 % to about 5 %, about 0.01 % to about 10 %, about 0.01 % to about 15 %, about 0.01 % to about 20 %, about 0.01 % to about 25 %, about 0.01 % to about 30 %, about 0.1 % to about 0.5 %, about 0.1 % to about 1 %, about 0.1 % to about 2 %, about 0.1 % to about 5 %, about 0.1 % to about 10 %, about 0.1 % to about 15 %, about 0.1 % to about 20 %, about 0.1 % to about 25 %, about 0.1 % to about 30 %, about 0.5 % to about 1 %, about 0.5 % to about 2 %, about 0.5 % to about 5 %, about 0.5 % to about 10 %, about 0.5 % to about 15 %, about 0.5 % to about 20 %, about 0.5 % to about 25 %, about 0.5 % to about 30 %, about 1 % to about 2 %, about 1 % to about 5 %, about 1 % to about 10 %, about 1 % to about 15 %, about 1 % to about 20 %, about 1 % to about 25 %, about 1 % to about 30 %, about 2 % to about 5 %, about 2 % to about 10 %, about 2 % to about 15 %, about 2 % to about 20 %, about 2 % to about 25 %, about 2 % to about 30 %, about 5 % to about 10 %, about 5 % to about 15 %, about 5 % to about 20 %, about 5 % to about 25 %, about 5 % to about 30 %, about 10 % to about 15 %, about 10 % to about 20 %, about 10 % to about 25 %, about 10 % to about 30 %, about 15 % to about 20 %, about 15 % to about 25 %, about 15 % to about 30 %, about 20 % to about 25 %, about 20 % to about 30 %, or about 25 % to about 30 %. In some instances, the data is recovered in the presence an error rate of about 0.001 %, about 0.01 %, about 0.1 %, about 0.5 %, about 1 %, about 2 %, about 5 %, about 10 %, about 15 %, about 20 %, about 25 %, or about 30 %. In some instances, the data is recovered in the presence an error rate of at least about 0.001 %, about 0.01 %, about 0.1 %, about 0.5 %, about 1 %, about 2 %, about 5 %, about 10 %, about 15 %, about 20 %, or about 25 %. In some instances, the data is recovered in the presence an error rate of at most about 0.01 %, about 0.1 %, about 0.5 %, about 1 %, about 2 %, about 5 %, about 10 %, about 15 %, about 20 %, about 25 %, or about 30 %.

[0181] In some instances, the decoding scheme comprises soft decoding. Soft decoding generally refers to decoding by considering a range of possible values (e.g., using probability estimates). As an example, sequencing carries quality for each base which can be considered during probability calculations. In such an example, each state comprises a final probability, which can be used in the outer decoder as, for example, log-likelihood if that outer decoder supports soft-decoding. Further, clustering and alignment can provide soft information on the alignment confidence. As an further example, an LDPC ECC comprises an iterative decoder. This provides possibilities to go back and forth between the inner and outer decoder in an iterative manner instead of a single pass. However, in some instances, this is accompanied by the cost of higher computing requirements.

[0182] The hashes of the present disclosure can allow verification of digital information during retrieval. In some cases, retrieving a digital information stored in a plurality of polynucleotides further comprises verifying at least the one or more objects. In some instances, the one or more objects are verified using a first one or more hashes in the plurality of pools. In some cases, retrieving a digital information stored in a plurality of polynucleotides further comprises verifying one or more pool items. In some instances, the one or more pool items are verified using a second one or more hashes in the plurality of pools. In some

examples, if an object is stored across more than one pool of the plurality of pools, more than one pool item is assembled into this object. In such examples, the first one or more hashes of one or more objects, the second one or more hashes of one or more pools items, or a combination thereof enables proper assembly verification.

[0183] Verifying hashes generally comprises generating hashes (e.g., cryptographic hashes). Verifying can further comprise comparing the generated hashes with the previously determined hashes. In some cases, the previously hashes and the new hashes are determined using the same hash function. In some instances, the hash function comprises a cryptographic hash function. In some cases, the hash function comprises MD-5, SHA-1, SHA-2, SHA-3, RIPEMD-160, Whirlpool, BLAKE, BLAKE2, BLAKE3, or a variation thereof. In some instances, the hash function comprises SHA-2. In some examples, SHA-2 comprises SHA-224, SHA-256, SHA-384, SHA-512, SHA-512/224, or SHA-512/256. In some cases, if the new and previous hashes match, the integrity of the item of information (e.g., an object) is verified. In some cases, if the new and previous hashes do not match, verification fails. In some instances, if verification fails, the integrity of the item of information is not verified. In some instances, if verification fails, the item of information has been modified and/or corrupted.

[0184] Retrieving digital information can comprise combining the information stored across pools items and/or the plurality of pools. In some cases, retrieving a digital information stored in a plurality of polynucleotides further comprises combining the digital information in the plurality of pools. In some instances, the data payload in the one or more pool items are combined. In some instances, the data payload in the one or more pool items across the plurality of pools are combined. In some instances, the combined data payloads comprise the digital information. In some cases, the retrieved digital information is further stored on a memory.

[0185] In some cases, the retrieved digital information is presented to a user. In some instances, the information is presented to a user on an interface. In some instances, the interface is an interface of an electronic device (e.g., personal electronic device). In some instances, the electronic device comprises an application configured to communicate with the systems described herein via a computer network to access the information.

[0186] The methods for retrieving digital information in DNA (or polynucleotides) can be carried out on a system. In some cases, such a system comprises an apparatus comprising one or more processing units, a memory, instructions, a sequencing device, or a combination thereof. In some instances, the memory is in communication with the one or more processing units. In some instances, the instructions are stored on the memory. In some instances the sequencing device in communication with the memory, the one or more processing units, or the combination thereof. In some cases, the one or more processing units and memory are distributed across one or more physical or logical locations.

[0187] In some instances, the memory is used to store digital information, polynucleotides sequences (e.g., partially or fully decoded sequences), or the combination thereof. In some instances, the memory is

used to store information related to the algorithms described herein (e.g., software code, parameters, executable instructions, etc.). In some examples, the memory can comprise any suitable memory described herein. In some examples, the memory can be configured according to embodiments described herein. In some examples, the sequencing device is configured to determining the plurality of nucleotide sequences using the methods described herein.

[0188] In some cases, the one or more processing units include any combination of central processing units (CPUs), graphical processing units (GPUs), single core processors, multi-core processors, processor clusters, application-specific integrated circuits (ASICs), programmable circuits such as Field Programmable Gate Arrays (FPGA), an AI-accelerator and variations thereof. In some cases, the one or more of the processing units comprise a Single Instruction Multiple Data (SIMD) or Single Program Multiple Data (SPMD) parallel architectures. As an example, the one or more processing units include one or more GPUs or CPUs that implement SIMD or SPMD. In some instances, an AI-accelerator comprise Google-TPU, Graphcore, Cerebras, SambaNova, or a combination thereof. In some embodiments, one or more of the processing units is implemented in software and/or firmware, in addition to hardware implementations. Software or firmware implementations of the processing units can include computer- or machine- executable instructions written in any suitable programming language to perform the various functions described herein. Software implementations of the one or more processing units can be stored in whole or part in the memory. Alternatively or additionally, the system can comprise one or more hardware logic components. For example, and without limitation, illustrative types of hardware logic components that can be used include Field-programmable Gate Arrays (FPGAs), Application-specific Integrated Circuits (ASICs), Application-specific Standard Products (ASSPs), System-on-a-chip systems (SOCs), Complex Programmable Logic Devices (CPLDs), etc. In some instances, decoding is run on compute-on-memory technologies, such as, but not limited to, UpMem.

[0189] In some instances, the one or more processing units is configured to perform one or more decoding steps. In some instances, the processing device is configured to perform one or more steps comprising: applying a decoding scheme to decode the digital information in the plurality of pools; verifying at least the one or more objects using a first one or more hashes in the plurality of pools; combining the digital information in the plurality of pools to retrieve the one or more objects; and storing the digital information on a memory. In some instances, the one or more processing units is configured to perform one or more steps comprising: apply an inner codec to the plurality of polynucleotides; or apply an ECC to the plurality of polynucleotides. In some instances, the inner codec transforms each of the plurality of polynucleotides into digital information. In some instances, the inner codec comprises a mixed decoding algorithm comprising a greedy algorithm and a maximum likelihood (ML) algorithm. In some instances, the output from an ECC are merged to generate an output comprising the digital information.

Certain definitions

[0190] Unless otherwise defined, all technical terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the present subject matter belongs.

[0191] Throughout this disclosure, numerical features are presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of any embodiments. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range to the tenth of the unit of the lower limit unless the context clearly dictates otherwise. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual values within that range, for example, 1.1, 2, 2.3, 5, and 5.9. This applies regardless of the breadth of the range. The upper and lower limits of these intervening ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention, unless the context clearly dictates otherwise.

[0192] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of any embodiment. As used herein, the singular forms “a,” “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof. As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0193] Reference throughout this specification to “some instances,” “further instances,” or “a particular instance,” means that a particular feature, structure, or characteristic described in connection with the instance is included in at least one instance. Thus, the appearances of the phrase “in some instances,” or “in further instances,” or “in a particular instance” in various places throughout this specification are not necessarily all referring to the same instance. Furthermore, the particular features, structures, or characteristics may be combined in any suitable manner in one or more instances.

[0194] Unless specifically stated or obvious from context, as used herein, the term “about” in reference to a number or range of numbers is understood to mean the stated number and numbers +/- 10% thereof, or 10% below the lower listed limit and 10% above the higher listed limit for the values listed for a range.

[0195] As used herein, the terms “preselected sequence”, “predefined sequence” or “predetermined sequence” are used interchangeably. The terms mean that the sequence of the polymer is known and chosen before synthesis or assembly of the polymer. In particular, various aspects of the invention are described herein primarily with regard to the preparation of nucleic acids molecules, the sequence of the

polynucleotide being known and chosen before the synthesis or assembly of the nucleic acid molecules.

[0196] As used herein, the term “hash” or “hashes” may generally refer to a string of fixed length that is outputted from a hash function. A hash function may generally comprise a function that receives an input of arbitrary length into an output with a fixed length. In some instances, the input may be one or more terms of a transaction or a contract, which may be passed through hash function to generate a hash. In some instances, the hash function may be deterministic, and it may be infeasible to reverse-engineer the input from the hashed output. The act of feeding an input into a hash function may be referred to as “hashing”.

[0197] Polynucleotide sequences described herein may be, unless stated otherwise, comprise DNA or RNA or an analog or derivative thereof. As used herein, the terms nucleic acids, polynucleotides, oligonucleotides, oligos, oligonucleic acids are used synonymously throughout to represent a polymer of nucleoside monomers. In some instances, nucleic acids are connected via phosphate or sulfur-containing linkages. Nucleic acids in some instances comprise DNA, RNA, non-canonical nucleic acids, unnatural nucleic acids, or other nucleoside. In some instances, nucleotides comprise non-canonical bases, sugars, or other moiety. In some instances, nucleotides comprise terminators which are configured to prevent extension reactions. In some instances, such terminators are removed before addition of subsequent nucleotides to the growing chain.

Computing system

[0198] Referring to **FIG. 12**, a block diagram is shown depicting an exemplary machine that includes a computer system **1200** (e.g., a processing or computing system) within which a set of instructions can execute for causing a device to perform or execute any one or more of the aspects and/or methodologies for static code scheduling of the present disclosure. The components in **FIG. 12** are examples only and do not limit the scope of use or functionality of any hardware, software, embedded logic component, or a combination of two or more such components implementing particular embodiments.

[0199] Computer system **1200** may include one or more processors **1201**, a memory **1203**, and a storage **1208** that communicate with each other, and with other components, via a bus **1240**. The bus **1240** may also link a display **1232**, one or more input devices **1233** (which may, for example, include a keypad, a keyboard, a mouse, a stylus, etc.), one or more output devices **1234**, one or more storage devices **1235**, and various tangible storage media **1236**. All of these elements may interface directly or via one or more interfaces or adaptors to the bus **1240**. For instance, the various tangible storage media **1236** can interface with the bus **1240** via storage medium interface **1226**. Computer system **1200** may have any suitable physical form, including but not limited to one or more integrated circuits (ICs), printed circuit boards (PCBs), mobile handheld devices (such as mobile telephones or PDAs), laptop or notebook computers, distributed computer systems, computing grids, or servers.

[0200] Computer system **1200** includes one or more processor(s) **1201** (e.g., central processing units

(CPUs), general purpose graphics processing units (GPGPUs), or quantum processing units (QPUs)) that carry out functions. Processor(s) **1201** optionally contains a cache memory unit **1202** for temporary local storage of instructions, data, or computer addresses. Processor(s) **1201** are configured to assist in execution of computer readable instructions. Computer system **1200** may provide functionality for the components depicted in **FIG. 12** as a result of the processor(s) **1201** executing non-transitory, processor-executable instructions embodied in one or more tangible computer-readable storage media, such as memory **1203**, storage **1208**, storage devices **1235**, and/or storage medium **1236**. The computer-readable media may store software that implements particular embodiments, and processor(s) **1201** may execute the software. Memory **1203** may read the software from one or more other computer-readable media (such as mass storage device(s) **1235**, **1236**) or from one or more other sources through a suitable interface, such as network interface **1220**. The software may cause processor(s) **1201** to carry out one or more processes or one or more steps of one or more processes described or illustrated herein. Carrying out such processes or steps may include defining data structures stored in memory **1203** and modifying the data structures as directed by the software.

[0201] The memory **1203** may include various components (e.g., machine readable media) including, but not limited to, a random access memory component (e.g., RAM **1204**) (e.g., static RAM (SRAM), dynamic RAM (DRAM), ferroelectric random access memory (FRAM), phase-change random access memory (PRAM), etc.), a read-only memory component (e.g., ROM **1205**), and any combinations thereof. ROM **1205** may act to communicate data and instructions unidirectionally to processor(s) **1201**, and RAM **1204** may act to communicate data and instructions bidirectionally with processor(s) **1201**. ROM **1205** and RAM **1204** may include any suitable tangible computer-readable media described below. In one example, a basic input/output system **1206** (BIOS), including basic routines that help to transfer information between elements within computer system **1200**, such as during start-up, may be stored in the memory **1203**.

[0202] Fixed storage **1208** is connected bidirectionally to processor(s) **1201**, optionally through storage control unit **1207**. Fixed storage **1208** provides additional data storage capacity and may also include any suitable tangible computer-readable media described herein. Storage **1208** may be used to store operating system **1209**, executable(s) **1210**, data **1211**, applications **1212** (application programs), and the like. Storage **1208** can also include an optical disk drive, a solid-state memory device (e.g., flash-based systems), or a combination of any of the above. Information in storage **1208** may, in appropriate cases, be incorporated as virtual memory in memory **1203**.

[0203] In one example, storage device(s) **1235** may be removably interfaced with computer system **1200** (e.g., via an external port connector (not shown)) via a storage device interface **1225**. Particularly, storage device(s) **1235** and an associated machine-readable medium may provide non-volatile and/or volatile storage of machine-readable instructions, data structures, program modules, and/or other data for the computer system **1200**. In one example, software may reside, completely or partially, within a machine-

readable medium on storage device(s) **1235**. In another example, software may reside, completely or partially, within processor(s) **1201**.

[0204] Bus **1240** connects a wide variety of subsystems. Herein, reference to a bus may encompass one or more digital signal lines serving a common function, where appropriate. Bus **1240** may be any of several types of bus structures including, but not limited to, a memory bus, a memory controller, a peripheral bus, a local bus, and any combinations thereof, using any of a variety of bus architectures. As an example and not by way of limitation, such architectures include an Industry Standard Architecture (ISA) bus, an Enhanced ISA (EISA) bus, a Micro Channel Architecture (MCA) bus, a Video Electronics Standards Association local bus (VLB), a Peripheral Component Interconnect (PCI) bus, a PCI-Express (PCI-X) bus, an Accelerated Graphics Port (AGP) bus, HyperTransport (HTX) bus, serial advanced technology attachment (SATA) bus, and any combinations thereof.

[0205] Computer system **1200** may also include an input device **1233**. In one example, a user of computer system **1200** may enter commands and/or other information into computer system **1200** via input device(s) **1233**. Examples of an input device(s) **1233** include, but are not limited to, an alpha-numeric input device (e.g., a keyboard), a pointing device (e.g., a mouse or touchpad), a touchpad, a touch screen, a multi-touch screen, a joystick, a stylus, a gamepad, an audio input device (e.g., a microphone, a voice response system, etc.), an optical scanner, a video or still image capture device (e.g., a camera), and any combinations thereof. In some embodiments, the input device is a Kinect, Leap Motion, or the like. Input device(s) **1233** may be interfaced to bus **1240** via any of a variety of input interfaces **1223** (e.g., input interface **1223**) including, but not limited to, serial, parallel, game port, USB, FIREWIRE, THUNDERBOLT, or any combination of the above.

[0206] In particular embodiments, when computer system **1200** is connected to network **1230**, computer system **1200** may communicate with other devices, specifically mobile devices and enterprise systems, distributed computing systems, cloud storage systems, cloud computing systems, and the like, connected to network **1230**. Communications to and from computer system **1200** may be sent through network interface **1220**. For example, network interface **1220** may receive incoming communications (such as requests or responses from other devices) in the form of one or more packets (such as Internet Protocol (IP) packets) from network **1230**, and computer system **1200** may store the incoming communications in memory **1203** for processing. Computer system **1200** may similarly store outgoing communications (such as requests or responses to other devices) in the form of one or more packets in memory **1203** and communicated to network **1230** from network interface **1220**. Processor(s) **1201** may access these communication packets stored in memory **1203** for processing.

[0207] Examples of the network interface **1220** include, but are not limited to, a network interface card, a modem, and any combination thereof. Examples of a network **1230** or network segment **1230** include, but are not limited to, a distributed computing system, a cloud computing system, a wide area network (WAN) (e.g., the Internet, an enterprise network), a local area network (LAN) (e.g., a network associated

with an office, a building, a campus or other relatively small geographic space), a telephone network, a direct connection between two computing devices, a peer-to-peer network, and any combinations thereof. A network, such as network 1230, may employ a wired and/or a wireless mode of communication. In general, any network topology may be used.

[0208] Information and data can be displayed through a display 1232. Examples of a display 1232 include, but are not limited to, a cathode ray tube (CRT), a liquid crystal display (LCD), a thin film transistor liquid crystal display (TFT-LCD), an organic liquid crystal display (OLED) such as a passive-matrix OLED (PMOLED) or active-matrix OLED (AMOLED) display, a plasma display, and any combinations thereof. The display 1232 can interface to the processor(s) 1201, memory 1203, and fixed storage 1208, as well as other devices, such as input device(s) 1233, via the bus 1240. The display 1232 is linked to the bus 1240 via a video interface 1222, and transport of data between the display 1232 and the bus 1240 can be controlled via the graphics control 1221. In some embodiments, the display is a video projector. In some embodiments, the display is a head-mounted display (HMD) such as a VR headset. In further embodiments, suitable VR headsets include, by way of non-limiting examples, HTC Vive, Oculus Rift, Samsung Gear VR, Microsoft HoloLens, Razer OSVR, FOVE VR, Zeiss VR One, Avegant Glyph, Freefly VR headset, and the like. In still further embodiments, the display is a combination of devices such as those disclosed herein.

[0209] In addition to a display 1232, computer system 1200 may include one or more other peripheral output devices 1234 including, but not limited to, an audio speaker, a printer, a storage device, and any combinations thereof. Such peripheral output devices may be connected to the bus 1240 via an output interface 1224. Examples of an output interface 1224 include, but are not limited to, a serial port, a parallel connection, a USB port, a FIREWIRE port, a THUNDERBOLT port, and any combinations thereof.

[0210] In addition or as an alternative, computer system 1200 may provide functionality as a result of logic hardwired or otherwise embodied in a circuit, which may operate in place of or together with software to execute one or more processes or one or more steps of one or more processes described or illustrated herein. Reference to software in this disclosure may encompass logic, and reference to logic may encompass software. Moreover, reference to a computer-readable medium may encompass a circuit (such as an IC) storing software for execution, a circuit embodying logic for execution, or both, where appropriate. The present disclosure encompasses any suitable combination of hardware, software, or both.

[0211] Those of skill in the art will appreciate that the various illustrative logical blocks, modules, circuits, and algorithm steps described in connection with the embodiments disclosed herein may be implemented as electronic hardware, computer software, or combinations of both. To clearly illustrate this interchangeability of hardware and software, various illustrative components, blocks, modules, circuits, and steps have been described above generally in terms of their functionality.

[0212] The various illustrative logical blocks, modules, and circuits described in connection with the

embodiments disclosed herein may be implemented or performed with a general purpose processor, a digital signal processor (DSP), an application specific integrated circuit (ASIC), a field programmable gate array (FPGA) or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to perform the functions described herein. A general purpose processor may be a microprocessor, but in the alternative, the processor may be any conventional processor, controller, microcontroller, or state machine. A processor may also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, a plurality of microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration.

[0213] The steps of a method or algorithm described in connection with the embodiments disclosed herein may be embodied directly in hardware, in a software module executed by one or more processor(s), or in a combination of the two. A software module may reside in RAM memory, flash memory, ROM memory, EPROM memory, EEPROM memory, registers, hard disk, a removable disk, a CD-ROM, or any other form of storage medium known in the art. An exemplary storage medium is coupled to the processor such the processor can read information from, and write information to, the storage medium. In the alternative, the storage medium may be integral to the processor. The processor and the storage medium may reside in an ASIC. The ASIC may reside in a user terminal. In the alternative, the processor and the storage medium may reside as discrete components in a user terminal.

[0214] In accordance with the description herein, suitable computing devices include, by way of non-limiting examples, server computers, desktop computers, laptop computers, notebook computers, sub-notebook computers, netbook computers, netpad computers, set-top computers, media streaming devices, handheld computers, Internet appliances, mobile smartphones, tablet computers, personal digital assistants, video game consoles, and vehicles. Those of skill in the art will also recognize that select televisions, video players, and digital music players with optional computer network connectivity are suitable for use in the system described herein. Suitable tablet computers, in various embodiments, include those with booklet, slate, and convertible configurations, known to those of skill in the art.

[0215] In some embodiments, the computing device includes an operating system configured to perform executable instructions. The operating system is, for example, software, including programs and data, which manages the device's hardware and provides services for execution of applications. Those of skill in the art will recognize that suitable server operating systems include, by way of non-limiting examples, FreeBSD, OpenBSD, NetBSD®, Linux, Apple® Mac OS X Server®, Oracle® Solaris®, Windows Server®, and Novell® NetWare®. Those of skill in the art will recognize that suitable personal computer operating systems include, by way of non-limiting examples, Microsoft® Windows®, Apple® Mac OS X®, UNIX®, and UNIX-like operating systems such as GNU/Linux®. In some embodiments, the operating system is provided by cloud computing. Those of skill in the art will also recognize that suitable mobile smartphone operating systems include, by way of non-limiting examples, Nokia® Symbian® OS, Apple®

iOS[®], Research In Motion[®] BlackBerry OS[®], Google[®] Android[®], Microsoft[®] Windows Phone[®] OS, Microsoft[®] Windows Mobile[®] OS, Linux[®], and Palm[®] WebOS[®]. Those of skill in the art will also recognize that suitable media streaming device operating systems include, by way of non-limiting examples, Apple TV[®], Roku[®], Boxee[®], Google TV[®], Google Chromecast[®], Amazon Fire[®], and Samsung[®] HomeSync[®]. Those of skill in the art will also recognize that suitable video game console operating systems include, by way of non-limiting examples, Sony[®] PS3[®], Sony[®] PS4[®], Microsoft[®] Xbox 360[®], Microsoft Xbox One, Nintendo[®] Wii[®], Nintendo[®] Wii U[®], and Ouya[®].

Non-transitory computer readable storage medium

[0216] In some embodiments, the platforms, systems, media, and methods disclosed herein include one or more non-transitory computer readable storage media encoded with a program including instructions executable by the operating system of an optionally networked computing device. In further embodiments, a computer readable storage medium is a tangible component of a computing device. In still further embodiments, a computer readable storage medium is optionally removable from a computing device. In some embodiments, a computer readable storage medium includes, by way of non-limiting examples, CD-ROMs, DVDs, flash memory devices, solid state memory, magnetic disk drives, magnetic tape drives, optical disk drives, distributed computing systems including cloud computing systems and services, and the like. In some cases, the program and instructions are permanently, substantially permanently, semi-permanently, or non-transitorily encoded on the media.

Computer program

[0217] In some embodiments, the platforms, systems, media, and methods disclosed herein include at least one computer program, or use of the same. A computer program includes a sequence of instructions, executable by one or more processor(s) of the computing device's CPU, written to perform a specified task. Computer readable instructions may be implemented as program modules, such as functions, objects, Application Programming Interfaces (APIs), computing data structures, and the like, that perform particular tasks or implement particular abstract data types. In light of the disclosure provided herein, those of skill in the art will recognize that a computer program may be written in various versions of various languages.

[0218] The functionality of the computer readable instructions may be combined or distributed as desired in various environments. In some embodiments, a computer program comprises one sequence of instructions. In some embodiments, a computer program comprises a plurality of sequences of instructions. In some embodiments, a computer program is provided from one location. In other embodiments, a computer program is provided from a plurality of locations. In various embodiments, a computer program includes one or more software modules. In various embodiments, a computer program includes, in part or in whole, one or more web applications, one or more mobile applications, one or more standalone applications, one or more web browser plug-ins, extensions, add-ins, or add-ons, or combinations thereof.

Web application

[0219] In some embodiments, a computer program includes a web application. In light of the disclosure provided herein, those of skill in the art will recognize that a web application, in various embodiments, utilizes one or more software frameworks and one or more database systems. In some embodiments, a web application is created upon a software framework such as Microsoft® .NET or Ruby on Rails (RoR). In some embodiments, a web application utilizes one or more database systems including, by way of non-limiting examples, relational, non-relational, object oriented, associative, XML, and document oriented database systems. In further embodiments, suitable relational database systems include, by way of non-limiting examples, Microsoft® SQL Server, MySQL™, and Oracle®. Those of skill in the art will also recognize that a web application, in various embodiments, is written in one or more versions of one or more languages. A web application may be written in one or more markup languages, presentation definition languages, client-side scripting languages, server-side coding languages, database query languages, or combinations thereof. In some embodiments, a web application is written to some extent in a markup language such as Hypertext Markup Language (HTML), Extensible Hypertext Markup Language (XHTML), or eXtensible Markup Language (XML). In some embodiments, a web application is written to some extent in a presentation definition language such as Cascading Style Sheets (CSS). In some embodiments, a web application is written to some extent in a client-side scripting language such as Asynchronous JavaScript and XML (AJAX), Flash® ActionScript, JavaScript, or Silverlight®. In some embodiments, a web application is written to some extent in a server-side coding language such as Active Server Pages (ASP), ColdFusion®, Perl, Java™, JavaServer Pages (JSP), Hypertext Preprocessor (PHP), Python™, Ruby, Tcl, Smalltalk, WebDNA®, or Groovy. In some embodiments, a web application is written to some extent in a database query language such as Structured Query Language (SQL). In some embodiments, a web application integrates enterprise server products such as IBM® Lotus Domino®. In some embodiments, a web application includes a media player element. In various further embodiments, a media player element utilizes one or more of many suitable multimedia technologies including, by way of non-limiting examples, Adobe® Flash®, HTML 5, Apple® QuickTime®, Microsoft® Silverlight®, Java™, and Unity®.

Mobile application

[0220] In some embodiments, a computer program includes a mobile application provided to a mobile computing device. In some embodiments, the mobile application is provided to a mobile computing device at the time it is manufactured. In other embodiments, the mobile application is provided to a mobile computing device via the computer network described herein.

[0221] In view of the disclosure provided herein, a mobile application is created by techniques known to those of skill in the art using hardware, languages, and development environments known to the art. Those of skill in the art will recognize that mobile applications are written in several languages. Suitable programming languages include, by way of non-limiting examples, C, C++, C#, Objective-C, Java™,

JavaScript, Pascal, Object Pascal, Python™, Ruby, VB.NET, WML, and XHTML/HTML with or without CSS, or combinations thereof.

[0222] Suitable mobile application development environments are available from several sources. Commercially available development environments include, by way of non-limiting examples, AirplaySDK, alcheMo, Appcelerator®, Celsius, Bedrock, Flash Lite, .NET Compact Framework, Rhomobile, and WorkLight Mobile Platform. Other development environments are available without cost including, by way of non-limiting examples, Lazarus, MobiFlex, MoSync, and Phonegap. Also, mobile device manufacturers distribute software developer kits including, by way of non-limiting examples, iPhone and iPad (iOS) SDK, Android™ SDK, BlackBerry® SDK, BREW SDK, Palm® OS SDK, Symbian SDK, webOS SDK, and Windows® Mobile SDK.

[0223] Those of skill in the art will recognize that several commercial forums are available for distribution of mobile applications including, by way of non-limiting examples, Apple® App Store, Google® Play, Chrome WebStore, BlackBerry® App World, App Store for Palm devices, App Catalog for webOS, Windows® Marketplace for Mobile, Ovi Store for Nokia® devices, Samsung® Apps, and Nintendo® DSi Shop.

Standalone application

[0224] In some embodiments, a computer program includes a standalone application, which is a program that is run as an independent computer process, not an add-on to an existing process, e.g., not a plug-in. Those of skill in the art will recognize that standalone applications are often compiled. A compiler is a computer program(s) that transforms source code written in a programming language into binary object code such as assembly language or machine code. Suitable compiled programming languages include, by way of non-limiting examples, C, C++, Objective-C, COBOL, Delphi, Eiffel, Java™, Lisp, Python™, Visual Basic, and VB .NET, or combinations thereof. Compilation is often performed, at least in part, to create an executable program. In some embodiments, a computer program includes one or more executable compiled applications.

Web browser plug-in

[0225] In some embodiments, the computer program includes a web browser plug-in (e.g., extension, etc.). In computing, a plug-in is one or more software components that add specific functionality to a larger software application. Makers of software applications support plug-ins to enable third-party developers to create abilities which extend an application, to support easily adding new features, and to reduce the size of an application. When supported, plug-ins enable customizing the functionality of a software application. For example, plug-ins are commonly used in web browsers to play video, generate interactivity, scan for viruses, and display particular file types. Those of skill in the art will be familiar with several web browser plug-ins including, Adobe® Flash® Player, Microsoft® Silverlight®, and Apple® QuickTime®. In some embodiments, the toolbar comprises one or more web browser extensions, add-ins,

or add-ons. In some embodiments, the toolbar comprises one or more explorer bars, tool bands, or desk bands.

[0226] In view of the disclosure provided herein, those of skill in the art will recognize that several plug-in frameworks are available that enable development of plug-ins in various programming languages, including, by way of non-limiting examples, C++, Delphi, Java™, PHP, Python™, and VB .NET, or combinations thereof.

[0227] Web browsers (also called Internet browsers) are software applications, designed for use with network-connected computing devices, for retrieving, presenting, and traversing information resources on the World Wide Web. Suitable web browsers include, by way of non-limiting examples, Microsoft® Internet Explorer®, Mozilla® Firefox®, Google® Chrome, Apple® Safari®, Opera Software® Opera®, and KDE Konqueror. In some embodiments, the web browser is a mobile web browser. Mobile web browsers (also called microbrowsers, mini-browsers, and wireless browsers) are designed for use on mobile computing devices including, by way of non-limiting examples, handheld computers, tablet computers, netbook computers, subnotebook computers, smartphones, music players, personal digital assistants (PDAs), and handheld video game systems. Suitable mobile web browsers include, by way of non-limiting examples, Google® Android® browser, RIM BlackBerry® Browser, Apple® Safari®, Palm® Blazer, Palm® WebOS® Browser, Mozilla® Firefox® for mobile, Microsoft® Internet Explorer® Mobile, Amazon® Kindle® Basic Web, Nokia® Browser, Opera Software® Opera® Mobile, and Sony® PSP™ browser.

Software modules

[0228] In some embodiments, the platforms, systems, media, and methods disclosed herein include software, server, and/or database modules, or use of the same. In view of the disclosure provided herein, software modules are created by techniques known to those of skill in the art using machines, software, and languages known to the art. The software modules disclosed herein are implemented in a multitude of ways. In various embodiments, a software module comprises a file, a section of code, a programming object, a programming structure, a distributed computing resource, a cloud computing resource, or combinations thereof. In further various embodiments, a software module comprises a plurality of files, a plurality of sections of code, a plurality of programming objects, a plurality of programming structures, a plurality of distributed computing resources, a plurality of cloud computing resources, or combinations thereof. In various embodiments, the one or more software modules comprise, by way of non-limiting examples, a web application, a mobile application, a standalone application, and a distributed or cloud computing application. In some embodiments, software modules are in one computer program or application. In other embodiments, software modules are in more than one computer program or application. In some embodiments, software modules are hosted on one machine. In other embodiments, software modules are hosted on more than one machine. In further embodiments, software modules are hosted on a distributed computing platform such as a cloud computing platform. In some embodiments,

software modules are hosted on one or more machines in one location. In other embodiments, software modules are hosted on one or more machines in more than one location.

Databases

[0229] In some embodiments, the platforms, systems, media, and methods disclosed herein include one or more databases, or use of the same. In view of the disclosure provided herein, those of skill in the art will recognize that many databases are suitable for storage and retrieval of information. In various embodiments, suitable databases include, by way of non-limiting examples, relational databases, non-relational databases, object oriented databases, object databases, entity-relationship model databases, associative databases, XML databases, document oriented databases, and graph databases. Further non-limiting examples include SQL, PostgreSQL, MySQL, Oracle, DB2, Sybase, and MongoDB. In some embodiments, a database is Internet-based. In further embodiments, a database is web-based. In still further embodiments, a database is cloud computing-based. In a particular embodiment, a database is a distributed database. In other embodiments, a database is based on one or more local computer storage devices.

EXAMPLES

[0230] The following illustrative examples are representative of embodiments of the software applications, systems, and methods described herein and are not meant to be limiting in any way.

Example 1 – Quality Control of Digital Information in DNA Post-Synthesis

[0231] Digital information is encoded in data polynucleotides using methods described herein (FIGs. 5-7). A voltage is applied to a synthesis surface and a current is measured to determine if there is a defect on the synthesis surface. About 100,000 data polynucleotides are synthesized using the methods described herein with continuous quality control by current sensing, optical imaging, and flow sensing. Post-synthesis of the polynucleotides, a distribution of the polynucleotide lengths and mass of the polynucleotides are estimated using absorbance and fluoresce measurements.

[0232] Quality control (QC) polynucleotides are synthesized on the synthesis surface with the data polynucleotides. The QC polynucleotides constitute about 1 % of the total polynucleotides on the surface (c.g., data polynucleotides plus QC polynucleotides). The QC polynucleotides comprise a first primer sequences, and are amplified based on the first primer sequence. The data polynucleotides comprise a second primer sequence that is different than the first primer sequence. The QC polynucleotides are fully sequenced. The QC polynucleotides are then aligned against a reference. The reference is a preselected sequence that is known. Aligning the QC polynucleotides against a reference generates a relative read count to determine the number of QC polynucleotides that have the same sequence as the reference sequence.

[0233] The QC polynucleotides are aligned against a reference to estimate an error rate in the polynucleotides. The error rate in the QC polynucleotides serve as a proxy for the error rate in the data

polynucleotides. The error rate is estimated to be less than 5 %. The QC polynucleotides are also aligned against a reference to estimate a synthesis uniformity in the data polynucleotides. The synthesis uniformity of the QC polynucleotides are analyzed across locations of the synthesis surface. The synthesis uniformity is estimated to be greater than 95 %.

[0234] A subset of the data polynucleotides comprising about 0.1 % of the data polynucleotides are also selected and amplified. The subset is randomly selected from across the synthesis surface, and are sequenced. The subset is partially decoded by an inner codec to determine an index, as shown in **FIG. 9**. The index is used to estimate a relative distribution of the subset of the plurality of data polynucleotides that are then arranged according to lanes and frames. Since the subset comprises about 0.1 % of 100,000 data polynucleotides, the relative distribution of the subset is centered around about every 100 decode indices. The relative distribution is used to estimate synthesis uniformity, which is estimated to be greater than 95 %.

[0235] A inner codec comprising a mixed greedy ML algorithm is further applied and a likelihood is generated. The likelihood is based on the number of steps required for decoding and the probability associated with each step in the algorithm. The likelihood is associated with an error rate in the data polynucleotides, where a high likelihood is associated with a low error rate, and a low likelihood is associated with a high error rate. The error rate is estimated to be less than 5 %.

[0236] The error and uniformity estimation from the QC polynucleotides and the error and uniformity estimation from the subset of data polynucleotides are then combined to determine a final pass or fail of the synthesized data polynucleotides. Since the error rate is less than 5% and the uniformity is greater than 95% from the quality control of both the QC polynucleotides and the subset of polynucleotides, the data polynucleotides pass the quality control.

[0237] While preferred embodiments of the present subject matter have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the present subject matter. It should be understood that various alternatives to the embodiments of the present subject matter described herein may be employed in practicing the present subject matter.

[0238] The present disclosure is further described by the following non-limiting items.

[0239] Item 1. A method for quality control (QC) of data polynucleotides, comprising:

- a. providing a plurality of QC polynucleotides on a surface, wherein the plurality of QC polynucleotides comprises a first primer sequence;
- b. amplifying the plurality of QC polynucleotides based on the first primer sequence;
- c. sequencing the plurality of QC polynucleotides; and

- d. aligning the plurality of QC polynucleotides against a reference to estimate an error rate in the data polynucleotides, a synthesis uniformity in the data polynucleotides, or a combination thereof for QC the data polynucleotides.

[0240] Item 2. The method of item 1, wherein the error rate, the synthesis uniformity, or a combination thereof is based at least in part on a relative read count of the plurality of QC polynucleotides.

[0241] Item 3. The method of items 1 or 2, wherein the plurality of QC polynucleotides is about or less 1% of the polynucleotides on the surface.

[0242] Item 4. The method of any previous item, wherein the plurality of QC polynucleotides are provided at a portion of the surface.

[0243] Item 5. The method of any previous item, wherein the plurality of QC polynucleotides are provided in uniformly on the surface.

[0244] Item 6. The method of any previous item, wherein the data polynucleotides comprise a second primer sequence.

[0245] Item 7. The method of item 6, wherein the first primer sequence is different than the second primer sequence.

[0246] Item 8. The method of item 6 or 7, wherein the first primer sequence and the second primer sequence are different lengths.

[0247] Item 9. The method of any previous items, wherein the quality control is performed after synthesis of the data polynucleotides.

[0248] Item 10. The method of item 9, where in the QC is performed prior to cleavage of the data polynucleotides from the surface.

[0249] Item 11. The method of any previous items, wherein each of the QC polynucleotides is about 50 to 200 nucleobases in length.

[0250] Item 12. The method of any previous items, wherein each of the data polynucleotides is about 100 to about 300 nucleobases in length.

[0251] Item 13. A method for quality control (QC) of data polynucleotides, comprising:

- a. selecting a subset of a plurality of data polynucleotides;
- b. applying an inner codec to the subset of the plurality of data polynucleotides, wherein the inner codec comprises probabilistic decoding; and
- c. estimating an error rate in the plurality of polynucleotides based at least in part on a likelihood associated with each decoded sequence in the subset of the plurality of data polynucleotides.

[0252] Item 14. The method of item 13, wherein a high likelihood is associated with a lower error rate.

[0253] Item 15. The method of items 13 or 14, wherein a low likelihood is associated with a higher error rate.

[0254] Item 16. The method of any one of items 13-15, further comprising decoding an index of the subset of the data polynucleotides.

[0255] Item 17. The method of item 16, wherein the index is decoded using the inner codec, an outer codec, or a combination thereof.

[0256] Item 18. The method of item 16, wherein the index is used to estimate a relative distribution of the subset of the plurality of data polynucleotides.

[0257] Item 19. The method of any one of items 13-18, wherein the QC is performed during synthesis of the polynucleotides, QC of stored polynucleotides, or a combination thereof.

[0258] Item 20. The method of any one of items 13-19, wherein the subset of the plurality of the data polynucleotides are selected at random.

[0259] Item 21. The method of any one of items 13-19, wherein the subset of the plurality of the data polynucleotides are selected based at least in part on their location on a surface.

[0260] Item 22. The method of any one of items 13-21, wherein the plurality of data polynucleotides comprises about 100,000 polynucleotides.

[0261] Item 23. The method of item 22, wherein the subset of the plurality of data polynucleotides is about 0.1 % of the plurality of data polynucleotides.

[0262] Item 24. The method of any previous item, wherein the method is used in conjunction with current sensing, optical imaging, flow sensing, size estimation, quality estimation, mass estimation, or any combination thereof.

[0263] Item 25. The method of item 24, wherein the current sensing comprises measuring a current of a chip or a section of the chip.

[0264] Item 26. The method of item 25, wherein the current is compared to a reference value.

[0265] Item 27. The method of item 26, wherein a difference between the current and the reference value is indicative of a chip failure, a deblocking failure, or a combination thereof.

[0266] Item 28. The method of any one of items 24-27, wherein the current sensing is performed before synthesis of the plurality of data polynucleotides.

[0267] Item 29. The method of item 28, wherein the current sensing is used to detect a chip defect, adjust polynucleotide synthesis locations on a chip, or a combination thereof.

[0268] Item 30. The method of any one of items 24-29, wherein mass estimation is performed using fluorescence.

[0269] Item 31. The method of item 30, wherein the fluorescence is used to detect a yield of the plurality of polynucleotides.

[0270] Item 32. The method of any one of items 24-31, wherein optical imaging comprises detecting a chip defect, non-uniformity, or a combination thereof.

[0271] Item 33. A method of performing QC of a plurality of cells on a surface, comprising:

- a. measuring a current of each cell in the plurality of cells on the surface;
- b. determining if one or more cells in the plurality of cells comprises a defect based at least in part on the current;
- c. synthesizing and/or storing polynucleotides at a second one or more cells in the plurality of cells, wherein the second one or more cells do not comprise the defect.

[0272] Item 34. The method of item 33, wherein the defect comprises a physical defect.

[0273] Item 35. The method of item 33 or 34, wherein the surface is a synthesis surface, a storage surface, or a combination thereof.

[0274] Item 36. The method of any one of items 33-35, further comprising blocking the one or more cells comprising the defect.

[0275] Item 37. The method of item 36, wherein blocking is performed by a protecting group on the surface.

[0276] Item 38. The method of item 37, wherein blocking is performed by a photolabile protecting group on the surface.

[0277] Item 39. The method of any one of items 36-38, wherein blocking is performed by selectively supplying energy to the one or more cells.

[0278] Item 40. The method of any one of items 36-39, wherein blocking is performed by a masking material.

[0279] Item 41. The method of any one of items 36-40, wherein blocking is performed by addressable control of each cell in the plurality of cells.

CLAIMS**WHAT IS CLAIMED IS:**

1. A method for quality control (QC) of data polynucleotides, comprising:
 - a. providing a plurality of QC polynucleotides on a surface, wherein the plurality of QC polynucleotides comprises a first primer sequence;
 - b. amplifying the plurality of QC polynucleotides based on the first primer sequence;
 - c. sequencing the plurality of QC polynucleotides; and
 - d. aligning the plurality of QC polynucleotides against a reference to estimate an error rate in the data polynucleotides, a synthesis uniformity in the data polynucleotides, or a combination thereof for QC the data polynucleotides.
2. The method of claim 1, wherein the error rate, the synthesis uniformity, or a combination thereof is based at least in part on a relative read count of the plurality of QC polynucleotides.
3. The method of claims 1 or 2, wherein the plurality of QC polynucleotides is about or less 1% of the polynucleotides on the surface.
4. The method of any previous claim, wherein the plurality of QC polynucleotides are provided at a portion of the surface.
5. The method of any previous claim, wherein the plurality of QC polynucleotides are provided in uniformly on the surface.
6. The method of any previous claim, wherein the data polynucleotides comprise a second primer sequence.
7. The method of claim 6, wherein the first primer sequence is different than the second primer sequence.
8. The method of claim 6 or 7, wherein the first primer sequence and the second primer sequence are different lengths.
9. The method of any previous claims, wherein the quality control is performed after synthesis of the data polynucleotides.

10. The method of claim 9, where in the QC is performed prior to cleavage of the data polynucleotides from the surface.
11. The method of any previous claims, wherein each of the QC polynucleotides is about 50 to 200 nucleobases in length.
12. The method of any previous claims, wherein each of the data polynucleotides is about 100 to about 300 nucleobases in length.
13. A method for quality control (QC) of data polynucleotides, comprising:
 - a. selecting a subset of a plurality of data polynucleotides;
 - b. applying an inner codec to the subset of the plurality of data polynucleotides, wherein the inner codec comprises probabilistic decoding; and
 - c. estimating an error rate in the plurality of polynucleotides based at least in part on a likelihood associated with each decoded sequence in the subset of the plurality of data polynucleotides.
14. The method of claim 13, wherein a high likelihood is associated with a lower error rate.
15. The method of claims 13 or 14, wherein a low likelihood is associated with a higher error rate.
16. The method of any one of claims 13-15, further comprising decoding an index of the subset of the data polynucleotides.
17. The method of claim 16, wherein the index is decoded using the inner codec, an outer codec, or a combination thereof.
18. The method of claim 16, wherein the index is used to estimate a relative distribution of the subset of the plurality of data polynucleotides.
19. The method of any one of claims 13-18, wherein the QC is performed during synthesis of the polynucleotides, QC of stored polynucleotides, or a combination thereof.
20. The method of any one of claims 13-19, wherein the subset of the plurality of the data polynucleotides are selected at random.