

SYSTEMS AND METHODS FOR DETERMINING NUCLEIC ACIDS

FIELD

The present invention generally relates to systems and methods for imaging or
5 determining nucleic acids, for instance, within cells. In some embodiments, the transcriptome of
a cell may be determined.

BACKGROUND

Single-molecule fluorescent in situ hybridization (smFISH) is a powerful method for
detecting individual mRNA molecules in cells. The high detection efficiency and large dynamic
10 range of this method provides exquisite detail into the expression state, spatial distribution within
cells and intact tissues, and variation among cells of individual mRNAs. Such approaches have
been essential to many recent insights into understanding gene regulation and expression. A
fundamental limitation of smFISH, however, is its low throughput, typically only a few genes at
a time. This low throughput is due to a lack of distinguishable probes with which to label cells
15 and the cost of producing large amounts of labeled probe required for high efficiency staining.
Thus, improvements in detecting mRNA molecules are needed.

SUMMARY

The present invention generally relates to systems and methods for imaging or
determining nucleic acids, for instance, within cells. In some embodiments, the transcriptome of
20 a cell may be determined. The subject matter of the present invention involves, in some cases,
interrelated products, alternative solutions to a particular problem, and/or a plurality of different
uses of one or more systems and/or articles.

In one aspect, the present invention is generally directed to a composition. According to
one set of embodiments, the composition comprises a plurality of nucleic acid probes, at least
25 some of which comprise a first portion comprising a target sequence and a plurality of read
sequences. In some cases, each comprises a first portion comprising a target sequence and a
plurality of read sequences. In some embodiments, the plurality of read sequences are
distributed on the plurality of nucleic acid probes so as to define an error-correcting code.

In another aspect, the present invention is generally directed to a method. In one set of
30 embodiments, the method includes acts of exposing a sample to a plurality of nucleic acid
probes; for each of the nucleic acid probes, determining binding of the nucleic acid probes within

the sample; creating codewords based on the binding of the nucleic acid probes; and for at least some of the codewords, matching the codeword to a valid codeword wherein, if no match is found, applying error correction to the codeword to form a valid codeword.

5 The method, in another set of embodiments, includes acts of exposing a sample to a plurality of nucleic acid probes, wherein the nucleic acid probes comprise a first portion comprising a target sequence and a second portion comprising one or more read sequences, and wherein at least some of the plurality of nucleic acid probes comprises distinguishable nucleic acid probes formed from combinatorial combination of one or more read sequences taken from a plurality of read sequences; and for each of the nucleic acid probes, determining binding of the
10 target sequences of the nucleic acid probes within the sample.

In yet another set of embodiments, the method includes acts of exposing a sample to a plurality of primary nucleic acid probes; exposing the plurality of primary nucleic acid probes to a sequence of secondary nucleic acid probes and determining fluorescence of each of the secondary nucleic acid probes within the sample; creating codewords based on fluorescence of
15 the secondary nucleic acid probes; and for at least some of the codewords, matching the codeword to a valid codeword wherein, if no match is found, applying error correction to the codeword to form a valid codeword.

In one set of embodiments, the method includes acts of exposing a plurality of primary nucleic acid probes to a sample; and exposing the plurality of nucleic acid probes to a sequence
20 of secondary nucleic acid probes and determining fluorescence of each of the secondary probes within the sample. In some embodiments, at least some of the plurality of secondary nucleic acid probes comprises distinguishable secondary nucleic acid probes formed from combinatorial combination of one or more read sequences taken from a plurality of read sequences.

In another set of embodiments, the method comprises acts of exposing a cell to a plurality
25 of nucleic acid probes, exposing the plurality of nucleic acid probes to a first secondary probe comprising a first signaling entity, determining the first signaling entity at a resolution better than 100 nm, inactivating the first signaling entity, exposing the plurality of nucleic acid probes to a second secondary probe comprising a second signaling entity, and determining the second signaling entity at a resolution better than 100 nm.

30 The method, in yet another set of embodiments, includes acts of exposing a cell to a plurality of nucleic acid probes, exposing the plurality of nucleic acid probes to a first secondary

probe comprising a first signaling entity, determining the first signaling entity using a super-resolution imaging technique, inactivating the first signaling entity, exposing the plurality of nucleic acid probes to a second secondary probe comprising a second signaling entity, and determining the second signaling entity using a super-resolution imaging technique.

5 In certain embodiments, the method comprises acts of associating a plurality of targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code space; associating a plurality of distinguishable read sequences with the plurality of codewords such that each distinguishable read sequence
10 represents a value of a position within the codewords; and forming a plurality of nucleic acid probes, each comprising a target sequence and one or more read sequences.

 In addition, in one set of embodiments, the method includes acts of associating a plurality of targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form
15 an error-checking and/or error-correcting code space; forming a plurality of nucleic acid probes each comprising a target sequence; and forming groups comprising the plurality of nucleic acid probes such that each group of nucleic acid probes corresponds to at least one common value of a position within the codewords.

 In another set of embodiments, the method includes acts of associating a plurality of
20 targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions that is less than the number of targets, and wherein each codeword is associated with a single target, associating a plurality of distinguishable read sequences with the plurality of codewords such that each distinguishable read sequence represents a value of a position within the codewords, and forming a plurality of nucleic acid
25 probes, each comprising a target sequence and one or more read sequences.

 The method, in still another set of embodiments, includes acts of exposing a plurality of nucleic acid probes to a cell, exposing the plurality of nucleic acid probes to a sequence of secondary probes and determining fluorescence of each of the secondary probes within the cell, and based on the sequence of fluorescence of each of the secondary probes, determining nucleic
30 acids within the cell.

In another set of embodiments, the method includes acts of associating a plurality of targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code space; forming a plurality of nucleic acid probes each comprising a target sequence; and forming groups comprising the plurality of nucleic acid probes such that each group of nucleic acid probes correspond to at least one common value of a position within the codewords.

In yet another set of embodiments, the method includes acts of exposing a cell to a plurality of nucleic acid probes, exposing the plurality of nucleic acid probes to a first secondary probe comprising a first signaling entity, determining the first signaling entity using a super-resolution imaging technique, inactivating the first signaling entity, exposing the plurality of nucleic acid probes to a second secondary probe comprising a second signaling entity, and determining the second signaling entity using a super-resolution imaging technique.

In still another set of embodiments, the method includes acts of exposing a cell to a plurality of nucleic acid probes, exposing the plurality of nucleic acid probes to a first secondary probe comprising a first signaling entity, determining the first signaling entity at a resolution better than 100 nm, inactivating the first signaling entity, exposing the plurality of nucleic acid probes to a second secondary probe comprising a second signaling entity, and determining the second signaling entity using a super-resolution imaging technique.

In still another set of embodiments, the method includes acts of associating a plurality of nucleic acid targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code; associating unique read sequences with each possible value of each position in the codewords, wherein the read sequences are taken from a set of orthogonal sequences, which have limited homology with one another and with the nucleic acid species in a sample; forming a plurality of primary nucleic acid probes each comprising a target sequence that uniquely binds to a nucleic acid target and one or more read sequences; forming a plurality of secondary nucleic acid probes comprising a signaling entity and a sequence that is complementary to one of the read sequences; exposing a sample to the primary nucleic acid probes such that the nucleic acid probes hybridize to the nucleic acid targets in the sample; exposing the primary nucleic acid probes in the sample to a secondary nucleic acid

probe such that the secondary nucleic acid probe hybridizes to the read sequence on at least some of the primary nucleic acid probes; imaging the sample; and repeating the exposing and imaging steps one or more times, using a different secondary nucleic acid probe for at least some of the repetitions.

5 The method, according to yet another set of embodiments, includes acts of associating a plurality of nucleic acid targets with a plurality of target sequences and a plurality of codewords wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code space; forming a plurality of nucleic acid probes comprising a signaling entity and a target sequence that uniquely binds to
10 one of the nucleic acid targets; grouping the nucleic acid probes into a plurality of probe pools, wherein each of the probe pools corresponds to a specific value of a unique position within the codewords; exposing a sample to one of the probe pools; imaging the sample; and repeating the exposing and imaging steps one or more times, using a different probe pool for at least some of the repetitions.

15 In another aspect, the present invention encompasses methods of making one or more of the embodiments described herein. In still another aspect, the present invention encompasses methods of using one or more of the embodiments described herein.

Other advantages and novel features of the present invention will become apparent from the following detailed description of various non-limiting embodiments of the invention when
20 considered in conjunction with the accompanying figures. In cases where the present specification and a document incorporated by reference include conflicting and/or inconsistent disclosure, the present specification shall control. If two or more documents incorporated by reference include conflicting and/or inconsistent disclosure with respect to each other, then the document having the later effective date shall control.

25 **BRIEF DESCRIPTION OF THE DRAWINGS**

Non-limiting embodiments of the present invention will be described by way of example with reference to the accompanying figures, which are schematic and are not intended to be drawn to scale. In the figures, each identical or nearly identical component illustrated is typically represented by a single numeral. For purposes of clarity, not every component is
30 labeled in every figure, nor is every component of each embodiment of the invention shown

where illustration is not necessary to allow those of ordinary skill in the art to understand the invention. In the figures:

Figs. 1A-1C illustrates an encoding scheme for nucleic acid probes, in certain embodiments of the invention;

5 Figs. 2A-2G illustrates determining of mRNAs in a cell, in some embodiments of the invention;

Figs. 3A-3B illustrate the determination of nucleic acids, in accordance with various embodiments of the invention; and

10 Figs. 4A-4B is a non-limiting example of multiple read sequences distributed in a population of different nucleic acid probes, in accordance with certain embodiments of the invention.

BRIEF DESCRIPTION OF THE SEQUENCES

SEQ ID NO: 1 is:

15 GTTGGCGACGAAAGCACTGCGATTGGAACCGTCCCAAGCGTTGCGCTTAATGGATC
ATCAATTTTGTCTCACTACGACGGTCAATCGCGCTGCATACTTGCGTCGGTCGGACA
AACGAGG;

SEQ ID NO: 2 is CGCAACGCTTGGGACGGTTCCAATCGGATC;

SEQ ID NO: 3 is CGAATGCTCTGGCCTCGAACGAACGATAGC;

SEQ ID NO: 4 is ACAAATCCGACCAGATCGGACGATCATGGG;

20 SEQ ID NO: 5 is CAAGTATGCAGCGCGATTGACCGTCTCGTT;

SEQ ID NO: 6 is TGCGTCGTCTGGCTAGCACGGCACGCAAAT;

SEQ ID NO: 7 is AAGTCGTACCCGATGCGCAGCAATTCACT;

SEQ ID NO: 8 is CGAAACATCGGCCACGGTCCCGTTGAACTT;

SEQ ID NO: 9 is ACGAATCCACCGTCCAGCGCGTCAAACAGA;

25 SEQ ID NO: 10 is CGCGAAATCCCCGTAACGAGCGTCCCCTTGC;

SEQ ID NO: 11 is GCATGAGTTGCCTGGCGTTGCGACGACTAA;

SEQ ID NO: 12 is CCGTCGTCTCCGGTCCACCGTTGCGCTTAC;

SEQ ID NO: 13 is GGCCAATGGCCCAGGTCCGTCACGCAATTT;

SEQ ID NO: 14 is TTGATCGAATCGGAGCGTAGCGGAATCTGC;

30 SEQ ID NO: 15 is CGCGCGGATCCGCTTGTCGGGAACGGATAC;

SEQ ID NO: 16 is GCCTCGATTACGACGGATGTAATTCGGCCG; and

SEQ ID NO: 17 is GCCCGTATTCCCGCTTGCGAGTAGGGCAAT.

DETAILED DESCRIPTION

The present invention generally relates to systems and methods for imaging or determining nucleic acids, for instance, within cells. In some embodiments, the transcriptome of a cell may be determined. Certain embodiments are directed to determining nucleic acids, such as mRNA, within cells at relatively high resolutions. In some embodiments, a plurality of nucleic acid probes may be applied to a sample, and their binding within the sample determined, e.g., using fluorescence, to determine locations of the nucleic acid probes within the sample. In some embodiments, codewords may be based on the binding of the plurality of nucleic acid probes, and in some cases, the codewords may define an error-correcting code to reduce or prevent misidentification of the nucleic acids. In certain cases, a relatively large number of different targets may be identified using a relatively small number of labels, e.g., by using various combinatorial approaches.

Two example approaches are now discussed. It should be understood, however, that these are presented by way of explanation and not limitation; other aspects and embodiments are discussed in further detail herein. In one example method, primary and secondary probes are used, where the primary probes encode “codewords” and bind to target nucleic acids in the sample, and the secondary probes are used to read out the codewords from the primary probes. In another example method, a plurality of different primary probes containing codewords are divided into as many separate pools as there are positions in the codewords, such that each primary probe pool corresponds to a certain value in a certain position of the codewords (e.g., a “one” in the first position as in “1001”).

The first example is now described with respect to Fig. 3A. As will be discussed in more detail below, in other embodiments, other configurations may be used as well. In this first example, a series of nucleic acid probes are used to determine nucleic acids within a cell or other sample, e.g., qualitatively or quantitatively. For example, nucleic acids may be identified as being present or absent, and/or the numbers or concentrations of certain nucleic acids may be determined within the cell or other sample. In some cases, the positions of the probes within the cell or other sample may be determined at relatively high resolutions, and in some cases, at resolutions better than the wavelength of visible light.

This example is generally directed to spatially detecting nucleic acids within a cell or other sample, e.g., at relatively high resolutions. For example, the nucleic acids may be mRNAs, or other nucleic acids described herein. In one set of embodiments, the nucleic acids within the cell may be determined by delivering or applying nucleic acid probes to the cell. In some cases, by using combinatorial approaches, a relatively large number of nucleic acids may be determined using a relatively small number of different labels on the nucleic acid probes. Thus, for example, a relatively small number of experiments may be used to determine a relatively large number of nucleic acids in a sample, e.g., due to simultaneous binding of the nucleic acid probes to different nucleic acids in the sample.

In one set of embodiments, a population of primary nucleic acid probes are applied to the cell (or other sample) that is able to bind nucleic acids suspected of being present within the cell. Afterwards, sequentially, secondary nucleic acid probes that can bind to or otherwise interact with some of the primary nucleic acids are added and determined, e.g., using imaging techniques such as STORM (stochastic optical reconstruction microscopy) or other imaging techniques. After imaging, the secondary nucleic acid probes are inactivated or removed, and a different secondary nucleic acid probe is added to the sample. This may be repeated multiple times with multiple different secondary nucleic acid probes. The pattern of binding of the various secondary nucleic acid probes may be used to determine the primary nucleic acid probes at locations within the cell or other sample, which can be used to determine mRNA or other nucleic acids that are present.

For instance, as is shown in Fig. 3A, a population of nucleic acids within a cell (represented here by nucleic acids 11, 12, and 13) may be exposed to a population of primary nucleic acid probes 20, including probes 21 and 22. The primary nucleic acid probes may contain, for instance, a target sequence that can recognize a nucleic acid (e.g., a sequence within nucleic acid 11). Probes 21 and 22 may contain the same or different targeting sequences, which may bind to or hybridize with the same or different nucleic acids. As an example, as is shown in Fig. 3A, probe 21 contains a first targeting sequence 25 which targets the probe to nucleic acid 11, while probe 22 contains a second targeting sequence 26, not identical to the first targeting sequence 25 and which targets the probe to nucleic acid 12. The target sequence may be substantially complementary to at least a portion of a target nucleic acid, and enough of the

target sequence may be present such that specific binding of the nucleic acid probe to the target nucleic acid can occur.

Primary nucleic acid probes 20 may also contain one or more “read” sequences. Two such read sequences are used in this example, although in other embodiments, there may be one, three, four, or more read sequences present within a primary nucleic acid probe. The read sequences may all independently be the same or different. In addition, in one set of embodiments, different nucleic acid probes may use one or more common read sequences. For example, more than one read sequence may be combinatorially combined on different nucleic acid probes, thereby producing a relatively large number of different nucleic acid probes that can be separately identified, even though only a relatively small number of read sequences are used. Thus, for example, in Fig. 3A, probe 21 contains read sequences 27 and 29, while probe 22 contains read sequences 27 and 28, where the two read sequences 27 are identical, and different from read sequences 28 and 29.

After primary nucleic acid probes 20 have been introduced to the sample and allowed to interact with nucleic acids 11, 12, and 13, one or more secondary nucleic acid probes 30 may be applied to the sample to determine the primary nucleic acid probes. The secondary nucleic acid probes may contain a recognition sequence able to recognize one of the read sequences present within the population of primary nucleic acid probes. For instance, the recognition sequence may be substantially complementary to at least a portion of the read sequence, such that the secondary nucleic acid probe is able to bind to or hybridize with corresponding primary nucleic acid probe. For instance, in this example, recognition sequence 35 is able to recognize read sequence 27. In addition, the secondary nucleic acid probes may contain one or more signaling entities 33. For example, a signaling entity may be a fluorescent entity attached to the probe, or a certain sequence of nucleic acids that can be determined in some fashion. More than one secondary sequence may be used, e.g., sequentially. For example, as shown in this figure, the initial secondary probe 30 may be removed (e.g., as discussed below) and a new secondary probe 31 may be added, containing recognition sequence 36 able to recognize read sequence 28 and one or more signaling entities 33. This may also be repeated multiple times, e.g., to determine read sequence 29 or other read sequences that may be present.

The location of the secondary nucleic acid probes 30, 31, etc. may be determined by determining signaling entity 33. For example, if the signaling entity is fluorescent, then

fluorescence microscopy can be used to determine the signaling entity. In some embodiments, imaging of a sample to determine the signaling entity may be used at relatively high resolutions, and in some cases, super-high resolution imaging techniques (e.g., resolutions better than the wavelength of visible light) may be used. Examples of super-high resolution imaging techniques include STORM, or other techniques as discussed herein. In some cases, e.g., with certain super-high resolution imaging techniques such as STORM, more than one image of the sample may be acquired.

More than one type of secondary nucleic acid probe may be applied to a cell or other sample. For example, a first secondary nucleic acid probe may be applied that can recognize a first read sequence, then it or its attached signaling entity may be inactivated or removed, and a second secondary nucleic acid probe may be applied that can recognize a second read sequence. This process may be repeated multiple times, each with a different secondary nucleic acid probe, e.g., to determine the read sequences that were present in the various primary nucleic acid probes. Thus, primary nucleic acids within the sample can be determined on the basis of the binding pattern of secondary nucleic acid probes.

For example, a first location within the cell or other sample may exhibit binding of a first secondary probe and a third secondary probe, but not the binding of a second or a fourth secondary probe, while a second location may exhibit a different pattern of binding of various secondary probes. The primary nucleic acid probe that the secondary probes are able to bind to or hybridize with may be determined by considering the pattern of binding of various secondary probes. For instance, referring to Fig. 3A, if a first secondary probe is able to determine read sequence 27, a second secondary probe is able to determine read sequence 28, and a third secondary probe is able to determine read sequence 29, then primary nucleic acid 25 may be determined through the binding of the first and third secondary probes (but not the second secondary probe), while primary nucleic acid 26 may be determined through the binding of the first and second secondary probes (but not the third secondary probe). Similarly, if it is known that first probe 21 contains target sequence 25 while second probe 22 contains target sequence 26, then nucleic acids 11 and 12 may also be determined within the sample, e.g., spatially, based on the binding pattern of the various secondary nucleic acid probes. In addition, it should be noted that due to the presence of more than one read sequence on the primary nucleic acid probes, even though first probe 21 and second probe 22 contains a common read sequence (read

sequence 27), these probes may be distinguished in the sample due to the different binding patterns of the various secondary nucleic acid probes.

In certain embodiments, this pattern of binding or hybridization of the secondary nucleic acid probes may be converted into a “codeword.” In this example, for instance, the codewords are “101” and “110” for first probe 21 and second probe 22, respectively, where a value of 1 represents binding and a value of 0 represents no binding. The codewords may also have longer lengths in other embodiments; only three probes are shown here for clarity purposes only. A codeword can be directly related to a specific target nucleic acid sequence of the primary nucleic acid probe. Accordingly, different primary nucleic acid probes may match certain codewords, which can then be used to identify the different targets of the primary nucleic acid probes based on the binding patterns of the secondary probes, even if in some cases, there is overlap in the read sequences of different secondary probes, e.g., as was shown in Fig. 3A. However, if no binding is evident (e.g., for nucleic acid 13), then the codeword would be “000” in this example.

The values in each codeword can also be assigned in different fashions in some embodiments. For example, a value of 0 could represent binding while a value of 1 represents no binding. Similarly, a value of 1 could represent binding of a secondary nucleic acid probe with one type of signaling entity while a value of 0 could represent binding of a secondary nucleic acid probe with another type of distinguishable signaling entity. These signaling entities could be distinguished, for example, via different colors of fluorescence. In some cases, values in codewords need not be confined to 0 and 1. The values could also be drawn from larger alphabets, such as ternary (e.g., 0, 1, and 2) or quaternary (e.g., 0, 1, 2, and 3) systems. Each different value could, for example, be represented by a different distinguishable signaling entity, including (in some cases) one value that may be represented by the absence of signal.

The codewords for each target may be assigned sequentially, or may be assigned at random. For instance, referring to Fig. 3A, a first nucleic acid target may be assigned to 101, while a second nucleic acid target may be assigned to 110. In addition, in some embodiments, the codewords may be assigned using an error-detection system or an error-correcting system, such as a Hamming system, a Golay code, or an extended Hamming system (or a SECDED system, i.e., single error correction, double error detection). Generally speaking, such systems can be used to identify where errors have occurred, and in some cases, such systems can also be used to correct the errors and determine what the correct codeword should have been. For

example, a codeword such as 001 may be detected as invalid and corrected using such a system to 101, e.g., if 001 is not previously assigned to a different target sequence. A variety of different error-correcting codes can be used, many of which have previously been developed for use within the computer industry; however, such error-correcting systems have not typically
5 been used within biological systems. Additional examples of such error-correcting codes are discussed in more detail below.

It should also be understood that all possible codewords in a code need not be used in some cases. For example, in some embodiments, codewords that are not used can serve as negative controls. Similarly, in some embodiments, some codewords can be left out because
10 they are more prone to errors in measurement than other codewords. For example, in some implementations, reading a codeword with more values of '1' might be more error-prone than reading a codeword with fewer values of '1.'

It should be understood that the above description is an example of one embodiment of the invention, and that primary and secondary nucleic acid probes are not necessary in all
15 embodiments. For example, in some embodiments, a series of nucleic acid probes containing signaling entities are used to determine nucleic acids within a cell or other sample, without necessarily requiring secondary probes.

For example, turning now to Fig. 3B, nucleic acids 11, 12, and 13 are exposed to different rounds of probes 21, 22, 23, 24, etc. in this example. These probes may each contain a
20 target sequence that can recognize a nucleic acid (e.g., a sequence within nucleic acid 11 or 12). These probes may each target the same nucleic acid, but different regions of the nucleic acid. In addition, some or all of the probes may contain one or more signaling entities, e.g., signaling entity 29 on probe 21. For example, the signaling entity may be a fluorescent entity attached to the probe, or a certain sequence of nucleic acids that can be determined in some fashion.

Probe 21 may be allowed to bind to nucleic acid 11 via target sequence 25. Such binding
25 can be determined by determining signaling entity 29. For example, if the signaling entity is fluorescent, then fluorescence microscopy can be used to determine the signaling entity, e.g., spatially within the cell or other sample. In some but not all embodiments, imaging of a sample to determine the signaling entity may be used at relatively high resolutions, and in some cases,
30 super-high resolution imaging techniques may be used. Other, different probes may be present as well; for instance, probe 22 containing target sequence 26 may bind to nucleic acid 12, and be

determined via signaling entity 29 within probe 22. These may occur, e.g., sequentially or simultaneously. Optionally, probes 21 and 22 may also be removed or inactivated, e.g., between application of different rounds of probes.

5 Next, a second round of probes (e.g., probe 23) is applied to the cell or other sample. In this example, probe 23 is able to bind to nucleic acid 11 via a targeting region, although there is no probe in the second round that is able to bind to nucleic acid 12. Binding of the probes is allowed to occur as discussed above, and determination of binding may occur via signaling entities. These signaling entities may be the same or different as from the first round of probes. This process may be repeated any number of times with different probes. For example, as is
10 shown in Fig. 3B, round 2 contains probes able to bind to nucleic acid 11, while round 3 contains probes able to bind to nucleic acid 12.

 In certain embodiments, each round of binding or hybridization of nucleic acid probes may be converted into a “codeword,” e.g., using only first probe 21 and second probe 22 the codewords 101 or 110 could be formed, where 1 represents binding and 0 represents no binding
15 and the first position corresponds to the binding of probe 21 while the second position corresponds to the binding of probe 22. A codeword of 000 would represent no binding, e.g., as shown with nucleic acid 13 in this example. A codeword can be directly related to a specific target nucleic acid sequence of the nucleic acid probes, by designing appropriate nucleic acid probes. Thus, for example, 110 may correspond to a first target nucleic acid 12 (e.g., the first
20 and second round of nucleic acid probes containing probes able to target nucleic acid 11, and these probes may target the same or different regions of nucleic acid 11) while 101 may correspond to a second target nucleic acid (e.g., the first and third round of nucleic acid probes containing probes able to target nucleic acid 12, and these probes may target the same or
25 different regions of nucleic acid 12). In addition, it should be noted that each round of probes may contain the same, or different signaling entities as other probes in the same round, and/or other probes in different rounds. For instance, in one set of embodiments, only one signaling entity is used in all of the rounds of probes.

 Similar to the above, the codewords for each target may be assigned sequentially, or may be assigned at random. The codewords may be assigned within a code space in some
30 embodiments using an error-detection or an error-correcting system, such as a Hamming system, a Golay code, or an extended Hamming system or a SECDED system (single error correction,

double error detection). Generally speaking, such error-correction systems can be used to identify where errors have occurred, and in some cases, such systems can also be used to correct the errors and determine what the correct codeword should have been.

Similar to the above, the values at each position in the codeword can be arbitrarily assigned in certain embodiments to binding or non-binding of probes that contain more than one distinguishable signaling entity.

In some cases, the nucleic acid probes may be formed into “pools” or groups of nucleic acids that share a common feature. For example, probes to all targets with codewords that contain a 1 in the first position, e.g. 110 and 101 but not 011, may comprise one pool while probes to all targets that contain a 1 in the second position, e.g. 110 and 011 but not 101, may comprise another pool. See also Fig. 1C. In some cases, a nucleic acid probe may be a member of more than one group or pool. Members of a nucleic acid pool may also contain features in addition to target sequences, read sequences, and or signaling entities that allow them to be distinguished from other groups. These features may be short nucleic acid sequences that are used for the amplification, production, or separation of these sequences. The nucleic acid probes of each group may be applied to a sample, e.g., sequentially, as discussed herein. Thus, in some aspects, the present invention is generally directed to systems and methods for determining nucleic acids within a cell or other sample. The sample may include a cell culture, a suspension of cells, a biological tissue, a biopsy, an organism, or the like. The sample may also be cell-free but nevertheless contain nucleic acids. If the sample contains a cell, the cell may be a human cell, or any other suitable cell, e.g., a mammalian cell, a fish cell, an insect cell, a plant cell, or the like. More than one cell may be present in some cases.

The nucleic acids to be determined may be, for example, DNA, RNA, or other nucleic acids that are present within a cell (or other sample). The nucleic acids may be endogenous to the cell, or added to the cell. For instance, the nucleic acid may be viral, or artificially created. In some cases, the nucleic acid to be determined may be expressed by the cell. The nucleic acid is RNA in some embodiments. The RNA may be coding and/or non-coding RNA. Non-limiting examples of RNA that may be studied within the cell include mRNA, siRNA, rRNA, miRNA, tRNA, lncRNA, snoRNAs, snRNAs, exRNAs, piRNAs, or the like.

In some cases, a significant portion of the nucleic acid within the cell may be studied. For instance, in some cases, enough of the RNA present within a cell may be determined so as to

produce a partial or complete transcriptome of the cell. In some cases, at least 4 types of mRNAs are determined within a cell, and in some cases, at least 3, at least 4, at least 7, at least 8, at least 12, at least 14, at least 15, at least 16, at least 22, at least 30, at least 31, at least 32, at least 50, at least 63, at least 64, at least 72, at least 75, at least 100, at least 127, at least 128, at least 140, at least 255, at least 256, at least 500, at least 1,000, at least 1,500, at least 2,000, at least 2,500, at least 3,000, at least 4,000, at least 5,000, at least 7,500, at least 10,000, at least 12,000, at least 15,000, at least 20,000, at least 25,000, at least 30,000, at least 40,000, at least 50,000, at least 75,000, or at least 100,000 types of mRNAs may be determined within a cell.

In some cases, the transcriptome of a cell may be determined. It should be understood that the transcriptome generally encompasses all RNA molecules produced within a cell, not just mRNA. Thus, for instance, the transcriptome may also include rRNA, tRNA, etc. In some embodiments, at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, or 100% of the transcriptome of a cell may be determined.

The determination of one or more nucleic acids within the cell or other sample may be qualitative and/or quantitative. In addition, the determination may also be spatial, e.g., the position of the nucleic acid within the cell or other sample may be determined in two or three dimensions. In some embodiments, the positions, number, and/or concentrations of nucleic acids within the cell (or other sample) may be determined.

As discussed herein, a variety of nucleic acid probes may be used to determine one or more nucleic acids within a cell or other sample. The probes may comprise nucleic acids (or entities that can hybridize to a nucleic acid, e.g., specifically) such as DNA, RNA, LNA (locked nucleic acids), PNA (peptide nucleic acids), or combinations thereof. In some cases, additional components may also be present within the nucleic acid probes, e.g., as discussed below. Any suitable method may be used to introduce nucleic acid probes into a cell.

For example, in some embodiments, the cell is fixed prior to introducing the nucleic acid probes, e.g., to preserve the positions of the nucleic acids within the cell. Techniques for fixing cells are known to those of ordinary skill in the art. As non-limiting examples, a cell may be fixed using chemicals such as formaldehyde, paraformaldehyde, glutaraldehyde, ethanol, methanol, acetone, acetic acid, or the like. In one embodiment, a cell may be fixed using Hepes-glutamic acid buffer-mediated organic solvent (HOPE).

The nucleic acid probes may be introduced into the cell (or other sample) using any suitable method. In some cases, the cell may be sufficiently permeabilized such that the nucleic acid probes may be introduced into the cell by flowing a fluid containing the nucleic acid probes around the cells. In some cases, the cells may be sufficiently permeabilized as part of a fixation
5 process; in other embodiments, cells may be permeabilized by exposure to certain chemicals such as ethanol, methanol, Triton, or the like. In addition, in some embodiments, techniques such as electroporation or microinjection may be used to introduce nucleic acid probes into a cell or other sample.

Certain aspects of the present invention are generally directed to nucleic acid probes that
10 are introduced into a cell (or other sample). The probes may comprise any of a variety of entities that can hybridize to a nucleic acid, typically by Watson-Crick base pairing, such as DNA, RNA, LNA, PNA, etc., depending on the application. The nucleic acid probe typically contains a target sequence that is able to bind to at least a portion of a target nucleic acid, in some cases specifically. When introduced into a cell or other system, the target system may be able to bind
15 to a specific target nucleic acid (e.g., an mRNA, or other nucleic acids as discussed herein). In some cases, the nucleic acid probes may be determined using signaling entities (e.g., as discussed below), and/or by using secondary nucleic acid probes able to bind to the nucleic acid probes (i.e., to primary nucleic acid probes). The determination of such nucleic acid probes is discussed in detail below.

20 In some cases, more than one type of (primary) nucleic acid probe may be applied to a sample, e.g., simultaneously. For example, there may be at least 2, at least 5, at least 10, at least 25, at least 50, at least 75, at least 100, at least 300, at least 1,000, at least 3,000, at least 10,000, or at least 30,000 distinguishable nucleic acid probes that are applied to a sample, e.g., simultaneously or sequentially.

25 The target sequence may be positioned anywhere within the nucleic acid probe (or primary nucleic acid probe). The target sequence may contain a region that is substantially complementary to a portion of a target nucleic acid. In some cases, the portions may be at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 92%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100%
30 complementary. In some cases, the target sequence may be at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 50, at least 60, at least 65, at least

75, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, at least 400, or at least 450 nucleotides in length. In some cases, the target sequence may be no more than 500, no more than 450, no more than 400, no more than 350, no more than 300, no more than 250, no more than 200, no more than 175, no more than 150, no more than 125, no more than 100, no more than 75, no more than 60, no more than 55, no more than 50, no more than 45, no more than 40, no more than 35, no more than 30, no more than 20, or no more than 10 nucleotides in length. Combinations of any of these are also possible, e.g., the target sequence may have a length of between 10 and 30 nucleotides, between 20 and 40 nucleotides, between 5 and 50 nucleotides, between 10 and 200 nucleotides, or between 25 and 35 nucleotides, between 10 and 300 nucleotides, etc. Typically, complementarity is determined on the basis of Watson-Crick nucleotide base pairing.

The target sequence of a (primary) nucleic acid probe may be determined with reference to a target nucleic acid suspected of being present within a cell or other sample. For example, a target nucleic acid to a protein may be determined using the protein's sequence, by determining the nucleic acids that are expressed to form the protein. In some cases, only a portion of the nucleic acids encoding the protein are used, e.g., having the lengths as discussed above. In addition, in some cases, more than one target sequence that can be used to identify a particular target may be used. For instance, multiple probes can be used, sequentially and/or simultaneously, that can bind to or hybridize to different regions of the same target.

Hybridization typically refers to an annealing process by which complementary single-stranded nucleic acids associate through Watson-Crick nucleotide base pairing (e.g., hydrogen bonding, guanine-cytosine and adenine-thymine) to form double-stranded nucleic acid.

In some embodiments, a nucleic acid probe, such as a primary nucleic acid probe, may also comprise one or more "read" sequences. However, it should be understood that read sequences are not necessary in all cases. In some embodiments, the nucleic acid probe may comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 or more, 20 or more, 32 or more, 40 or more, 50 or more, 64 or more, 75 or more, 100 or more, 128 or more read sequences. The read sequences may be positioned anywhere within the nucleic acid probe. If more than one read sequence is present, the read sequences may be positioned next to each other, and/or interspersed with other sequences.

The read sequences, if present, may be of any length. If more than one read sequence is used, the read sequences may independently have the same or different lengths. For instance, the read sequence may be at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 50, at least 60, at least 65, at least 75, at least 100, at least 125, at least 150, at least 175, at least 200, at least 250, at least 300, at least 350, at least 400, or at least 450 nucleotides in length. In some cases, the read sequence may be no more than 500, no more than 450, no more than 400, no more than 350, no more than 300, no more than 250, no more than 200, no more than 175, no more than 150, no more than 125, no more than 100, no more than 75, no more than 60, no more than 65, no more than 60, no more than 55, no more than 50, no more than 45, no more than 40, no more than 35, no more than 30, no more than 20, or no more than 10 nucleotides in length. Combinations of any of these are also possible, e.g., the read sequence may have a length of between 10 and 30 nucleotides, between 20 and 40 nucleotides, between 5 and 50 nucleotides, between 10 and 200 nucleotides, or between 25 and 35 nucleotides, between 10 and 300 nucleotides, etc.

The read sequence may be arbitrary or random in some embodiments. In certain cases, the read sequences are chosen so as to reduce or minimize homology with other components of the cell or other sample, e.g., such that the read sequences do not themselves bind to or hybridize with other nucleic acids suspected of being within the cell or other sample. In some cases, the homology may be less than 10%, less than 8%, less than 7%, less than 6%, less than 5%, less than 4%, less than 3%, less than 2%, or less than 1%. In some cases, there may be a homology of less than 20 basepairs, less than 18 basepairs, less than 15 basepairs, less than 14 basepairs, less than 13 basepairs, less than 12 basepairs, less than 11 basepairs, or less than 10 basepairs. In some cases, the basepairs are sequential.

In one set of embodiments, a population of nucleic acid probes may contain a certain number of read sequences, which may be less than the number of targets of the nucleic acid probes in some cases. Those of ordinary skill in the art will be aware that if there is one signaling entity and n read sequences, then in general $2^n - 1$ different nucleic acid targets may be uniquely identified. However, not all possible combinations need be used. For instance, a population of nucleic acid probes may target 12 different nucleic acid sequences, yet contain no more than 8 read sequences. As another example, a population of nucleic acids may target 140 different nucleic acid species, yet contain no more than 16 read sequences. Different nucleic acid

sequence targets may be separately identified by using different combinations of read sequences within each probe. For instance, each probe may contain 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, etc. or more read sequences. In some cases, a population of nucleic acid probes may each contain the same number of read sequences, although in other cases, there may be different
5 numbers of read sequences present on the various probes.

As a non-limiting example, a first nucleic acid probe may contain a first target sequence, a first read sequence, and a second read sequence, while a second, different nucleic acid probe may contain a second target sequence, the same first read sequence, but a third read sequence instead of the second read sequence. Such probes may thereby be distinguished by determining
10 the various read sequences present or associated with a given probe or location, as discussed herein.

In some embodiments, the nucleic acid probe may contain a signaling entity. It should be understood that signaling entities are not required in all cases, however; for instance, the nucleic acid probe may be determined using secondary nucleic acid probes in some embodiments, as is
15 discussed in additional detail below. Examples of signaling entities that can be used are also discussed in more detail below.

Other components may also be present within a nucleic acid probe as well. For example, in one set of embodiments, one or more primer sequences may be present, e.g., to allow for enzymatic amplification of probes. Those of ordinary skill in the art will be aware of primer
20 sequences suitable for applications such as amplification (e.g., using PCR or other suitable techniques). Many such primer sequences are available commercially. Other examples of sequences that may be present within a primary nucleic acid probe include, but are not limited to promoter sequences, operons, identification sequences, nonsense sequences, or the like.

Typically, a primer is a single-stranded or partially double-stranded nucleic acid (e.g.,
25 DNA) that serves as a starting point for nucleic acid synthesis, allowing polymerase enzymes such as nucleic acid polymerase to extend the primer and replicate the complementary strand. A primer is (e.g., is designed to be) complementary to and to hybridize to a target nucleic acid. In some embodiments, a primer is a synthetic primer. In some embodiments, a primer is a non-naturally-occurring primer. A primer typically has a length of 10 to 50 nucleotides. For
30 example, a primer may have a length of 10 to 40, 10 to 30, 10 to 20, 25 to 50, 15 to 40, 15 to 30,

20 to 50, 20 to 40, or 20 to 30 nucleotides. In some embodiments, a primer has a length of 18 to 24 nucleotides.

In addition, the components of the nucleic acid probe may be arranged in any suitable order. For instance, in one embodiment, the components may be arranged in a nucleic acid probe as: primer—read sequences—targeting sequence—read sequences—reverse primer. The “read sequences” in this structure may each contain any number (including 0) of read sequences, so long as at least one read sequence is present in the probe. Non-limiting example structures include primer—targeting sequence—read sequences—reverse primer, primer—read sequences—targeting sequence—reverse primer, targeting sequence—primer—targeting sequence—read sequences—reverse primer, targeting sequence—primer—read sequences—targeting sequence—reverse primer, primer—target sequence—read sequences—targeting sequence—reverse primer, targeting sequence—primer—read sequence—reverse primer, targeting sequence—read sequence—primer, read sequence—targeting sequence—primer, read sequence—primer—targeting sequence—reverse primer, etc. In addition, the reverse primer is optional in some embodiments, including in all of the above-described examples.

After introduction of the nucleic acid probes into a cell or other sample, the nucleic acid probes may be directly determined by determining signaling entities (if present), and/or the nucleic acid probes may be determined by using one or more secondary nucleic acid probes, in accordance with certain aspects of the invention. As mentioned, in some cases, the determination may be spatial, e.g., in two or three dimensions. In addition, in some cases, the determination may be quantitative, e.g., the amount or concentration of a primary nucleic acid probe (and of a target nucleic acid) may be determined. Additionally, the secondary probes may comprise any of a variety of entities able to hybridize a nucleic acid, e.g., DNA, RNA, LNA, and/or PNA, etc., depending on the application. Signaling entities are discussed in more detail below.

A secondary nucleic acid probe may contain a recognition sequence able to bind to or hybridize with a read sequence of a primary nucleic acid probe. In some cases, the binding is specific, or the binding may be such that a recognition sequence preferentially binds to or hybridizes with only one of the read sequences that are present. The secondary nucleic acid probe may also contain one or more signaling entities. If more than one secondary nucleic acid probe is used, the signaling entities may be the same or different.

The recognition sequences may be of any length, and multiple recognition sequences may be of the same or different lengths. If more than one recognition sequence is used, the recognition sequences may independently have the same or different lengths. For instance, the recognition sequence may be at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, or at least 50 nucleotides in length. In some cases, the recognition sequence may be no more than 75, no more than 60, no more than 65, no more than 60, no more than 55, no more than 50, no more than 45, no more than 40, no more than 35, no more than 30, no more than 20, or no more than 10 nucleotides in length. Combinations of any of these are also possible, e.g., the recognition sequence may have a length of between 10 and 30, between 20 and 40, or between 25 and 35 nucleotides, etc. In one embodiment, the recognition sequence is of the same length as the read sequence. In addition, in some cases, the recognition sequence may be at least 50%, at least 60%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 92%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or at least 100% complementary to a read sequence of the primary nucleic acid probe.

As mentioned, in some cases, the secondary nucleic acid probe may comprise one or more signaling entities. Examples of signaling entities are discussed in more detail below.

As discussed, in certain aspects of the invention, nucleic acid probes are used that contain various “read sequences.” For example, a population of primary nucleic acid probes may contain certain “read sequences” which can bind certain of the secondary nucleic acid probes, and the locations of the primary nucleic acid probes are determined within the sample using secondary nucleic acid probes, e.g., which comprise a signaling entity. As mentioned, in some cases, a population of read sequences may be combined in various combinations to produce different nucleic acid probes, e.g., such that a relatively small number of read sequences may be used to produce a relatively large number of different nucleic acid probes.

Thus, in some cases, a population of primary nucleic acid probes (or other nucleic acid probes) may each contain a certain number of read sequences, some of which are shared between different primary nucleic acid probes such that the total population of primary nucleic acid probes may contain a certain number of read sequences. A population of nucleic acid probes may have any suitable number of read sequences. For example, a population of primary nucleic acid probes may have 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 etc. read sequences. More than 20 are also possible in some embodiments. In addition, in some cases, a

population of nucleic acid probes may, in total, have 1 or more, 2 or more, 3 or more, 4 or more, 5 or more, 6 or more, 7 or more, 8 or more, 9 or more, 10 or more, 11 or more, 12 or more, 13 or more, 14 or more, 15 or more, 16 or more, 20 or more, 24 or more, 32 or more, 40 or more, 50 or more, 60 or more, 64 or more, 100 or more, 128 or more, etc. of possible read sequences present, although some or all of the probes may each contain more than one read sequence, as discussed herein. In addition, in some embodiments, the population of nucleic acid probes may have no more than 100, no more than 80, no more than 64, no more than 60, no more than 50, no more than 40, no more than 32, no more than 24, no more than 20, no more than 16, no more than 15, no more than 14, no more than 13, no more than 12, no more than 11, no more than 10, no more than 9, no more than 8, no more than 7, no more than 6, no more than 5, no more than 4, no more than 3, or no more than two read sequences present. Combinations of any of these are also possible, e.g., a population of nucleic acid probes may comprise between 10 and 15 read sequences in total.

As a non-limiting example of an approach to combinatorially producing a relatively large number of nucleic acid probes from a relatively small number of read sequences, in a population of 6 different types of nucleic acid probes, each comprising one or more read sequences, the total number of read sequences within the population may be no greater than 4. It should be understood that although 4 read sequences are used in this example for ease of explanation, in other embodiments, larger numbers of nucleic acid probes may be realized, for example, using 5, 8, 10, 16, 32, etc. or more read sequences, or any other suitable number of read sequences described herein, depending on the application. Referring now to Fig. 4A, if each of the primary nucleic acid probes contains two different read sequences, then by using 4 such read sequences (A, B, C, and D), up to 6 probes may be separately identified. It should be noted that in this example, the ordering of read sequences on a nucleic acid probe is not essential, i.e., "AB" and "BA" may be treated as being synonymous (although in other embodiments, the ordering of read sequences may be essential and "AB" and "BA" may not necessarily be synonymous). Similarly, if 5 read sequences are used (A, B, C, D, and E) in the population of primary nucleic acid probes, up to 10 probes may be separately identified, as is shown in Fig. 4B. One of ordinary skill in the art would understand that, for k read sequences in a population with n read sequences on each probe, up to $\binom{n}{k}$ different probes may be produced, assuming that the ordering of read sequences is not essential; fewer than this number of different probes may also be used in

certain embodiments. In addition, it should also be understood that the number of read sequences on each probe need not be identical in some embodiments. For instance example, some probes may contain 2 read sequences while other probes may contain 3 read sequences.

5 In some aspects, the read sequences and/or the pattern of binding of nucleic acid probes within a sample may be used to define an error-detecting and/or an error-correcting code, for example, to reduce or prevent misidentification or errors of the nucleic acids, e.g., as was discussed with reference to Fig. 3. Thus, for example, if binding is indicated (e.g., as determined using a signaling entity), then the location may be identified with a “1”; conversely, if no binding is indicated, then the location may be identified with a “0” (or vice versa, in some cases).
10 Multiple rounds of binding determinations, e.g., using different nucleic acid probes, can then be used to create a “codeword,” e.g., for that spatial location. In some embodiments, the codeword may be subjected to error detection and/or correction. For instance, the codewords may be organized such that, if no match is found for a given set of read sequences or binding pattern of nucleic acid probes, then the match may be identified as an error, and optionally, error correction
15 may be applied sequences to determine the correct target for the nucleic acid probes. In some cases, the codewords may have fewer “letters” or positions than the total number of nucleic acids encoded by the codewords, e.g. where each codeword encodes a different nucleic acid.

Such error-detecting and/or the error-correction code may take a variety of forms. A variety of such codes have previously been developed in other contexts such as the
20 telecommunications industry, such as Golay codes or Hamming codes. In one set of embodiments, the read sequences or binding patterns of the nucleic acid probes are assigned such that not every possible combination is assigned.

For example, if 4 read sequences are possible and a primary nucleic acid probe contains 2 read sequences, then up to 6 primary nucleic acid probes could be identified; but the number of
25 primary nucleic acid probes used may be less than 6. Similarly, for k read sequences in a population with n read sequences on each primary nucleic acid probe, up to $\binom{n}{k}$ different probes may be produced, but the number of primary nucleic acid probes that are used may be any number less than $\binom{n}{k}$. In addition, these may be randomly assigned, or assigned in specific ways to increase the ability to detect and/or correct errors.

As another example, if multiple rounds of nucleic acid probes are used, the number of rounds may be arbitrarily chosen. Up to 2^n different probes may be possible for n rounds of probes, but the number of nucleic acid probes that are actually used may be any number less than 2^n . In addition, these may be randomly assigned, or assigned in specific ways to increase the ability to detect and/or correct errors.

For example, in one set of embodiments, the codewords or nucleic acid probes may be assigned within a code space such that the assignments are separated by a Hamming distance, which measures the number of incorrect “reads” in a given pattern that cause the nucleic acid probe to be misinterpreted as a different valid nucleic acid probe. In certain cases, the Hamming distance may be at least 2, at least 3, at least 4, at least 5, at least 6, or the like. In addition, in one set of embodiments, the assignments may be formed as a Hamming code, for instance, a Hamming(7, 4) code, a Hamming(15, 11) code, a Hamming(31, 26) code, a Hamming(63, 57) code, a Hamming(127, 120) code, etc. In another set of embodiments, the assignments may form a SECDED code, e.g., a SECDED(8,4) code, a SECDED(16,4) code, a SCEDED(16, 11) code, a SCEDED(22, 16) code, a SCEDED(39, 32) code, a SCEDED(72, 64) code, etc. In yet another set of embodiments, the assignments may form an extended binary Golay code, a perfect binary Golay code, or a ternary Golay code. In another set of embodiments, the assignments may represent a subset of the possible values taken from any of the codes described above.

Accordingly, in some embodiments, once the codeword is determined (e.g., as discussed herein), the codeword may be compared to the known nucleic acid probe codewords. If a match is found, then the nucleic acid probe can be identified or determined. If no match is found, then an error in the reading of the codeword may be identified. In some cases, error correction can also be applied to determine the correct codeword, and thus the correct nucleic acid probe. In some cases, the codewords may be selected such that, assuming that there is only one error present, only one possible correct codeword is available, and thus, only one of the nucleic acid probes is possible. In some cases, this may also be generalized to larger codeword spacings or Hamming distances; for instance, the codewords may be selected such that if two, three, or four errors are present (or more in some cases), only one possible correct codeword is available, and thus, only one of the nucleic acid probes is possible.

The error-correcting code may be a binary error-correcting code, or it may be based on other numbering systems, e.g., ternary or quaternary error-correcting codes. For instance, in one

set of embodiments, more than one type of signaling entity may be used and assigned to different numbers within the error-correcting code. Thus, as a non-limiting example, a first signaling entity (or more than one signaling entity, in some cases) may be assigned as “1” and a second signaling entity (or more than one signaling entity, in some cases) may be assigned as “2” (with “0” indicating no signaling entity present), and the codewords distributed to define a ternary error-correcting code. Similarly, a third signaling entity may additionally be assigned as “3” to make a quaternary error-correcting code, etc.

As discussed above, in certain aspects, signaling entities are determined, e.g., to determine nucleic acid probes and/or to create codewords. In some cases, signaling entities within a sample may be determined, e.g., spatially, using a variety of techniques. In some embodiments, the signaling entities may be fluorescent, and techniques for determining fluorescence within a sample, such as fluorescence microscopy or confocal microscopy, may be used to spatially identify the positions of signaling entities within a cell. In some cases, the positions of entities within the sample may be determined in two or even three dimensions.

In some embodiments, the spatial positions of the entities (and thus, nucleic acid probes that the entities may be associated with) may be determined at relatively high resolutions. For instance, the positions may be determined at spatial resolutions of better than about 100 micrometers, better than about 30 micrometers, better than about 10 micrometers, better than about 3 micrometers, better than about 1 micrometer, better than about 800 nm, better than about 600 nm, better than about 500 nm, better than about 400 nm, better than about 300 nm, better than about 200 nm, better than about 100 nm, better than about 90 nm, better than about 80 nm, better than about 70 nm, better than about 60 nm, better than about 50 nm, better than about 40 nm, better than about 30 nm, better than about 20 nm, or better than about 10 nm, etc.

There are a variety of techniques able to determine or image the spatial positions of entities optically, e.g., using fluorescence microscopy. In some cases, the spatial positions may be determined at super resolutions, or at resolutions better than the wavelength of light. Non-limiting examples include STORM (stochastic optical reconstruction microscopy), STED (stimulated emission depletion microscopy), NSOM (Near-field Scanning Optical Microscopy), 4Pi microscopy, SIM (Structured Illumination Microscopy), SMI (Spatially Modulated Illumination) microscopy, RESOLFT (Reversible Saturable Optically Linear Fluorescence Transition Microscopy), GSD (Ground State Depletion Microscopy), SSIM (Saturated

Structured-Illumination Microscopy), SPDM (Spectral Precision Distance Microscopy), Photo-Activated Localization Microscopy (PALM), Fluorescence Photoactivation Localization Microscopy (FPALM), LIMON (3D Light Microscopical Nanosizing Microscopy), Super-resolution optical fluctuation imaging (SOFI), or the like. See, e.g., U.S. Pat. No. 7,838,302, issued November 23, 2010, entitled “Sub-Diffraction Limit Image Resolution and Other Imaging Techniques,” by Zhuang, *et al.*; U.S. Pat. No. 8,564,792, issued October 22, 2013, entitled “Sub-diffraction Limit Image Resolution in Three Dimensions,” by Zhuang, *et al.*; or Int. Pat. Apl. Pub. No. WO 2013/090360, published June 20, 2013, entitled “High Resolution Dual-Objective Microscopy,” by Zhuang, *et al.*, each incorporated herein by reference in their entireties.

In addition, the signaling entity may be inactivated in some cases. For example, in some embodiments, a first secondary nucleic acid probe containing a signaling entity may be applied to a sample that can recognize a first read sequence, then the first secondary nucleic acid probe can be inactivated before a second secondary nucleic acid probe is applied to the sample. If multiple signaling entities are used, the same or different techniques may be used to inactivate the signaling entities, and some or all of the multiple signaling entities may be inactivated, e.g., sequentially or simultaneously.

Inactivation may be caused by removal of the signaling entity (e.g., from the sample, or from the nucleic acid probe, etc.), and/or by chemically altering the signaling entity in some fashion, e.g., by photobleaching the signaling entity, bleaching or chemically altering the structure of the signaling entity, etc.). For instance, in one set of embodiments, a fluorescent signaling entity may be inactivated by chemical or optical techniques such as oxidation, photobleaching, chemically bleaching, stringent washing or enzymatic digestion or reaction by exposure to an enzyme, dissociating the signaling entity from other components (e.g., a probe), chemical reaction of the signaling entity (e.g., to a reactant able to alter the structure of the signaling entity) or the like.

In some embodiments, various nucleic acid probes (including primary and/or secondary nucleic acid probes) may include one or more signaling entities. If more than one nucleic acid probe is used, the signaling entities may each be the same or different. In certain embodiments, a signaling entity is any entity able to emit light. For instance, in one embodiment, the signaling entity is fluorescent. In other embodiments, the signaling entity may be phosphorescent, radioactive, absorptive, etc. In some cases, the signaling entity is any entity that can be

determined within a sample at relatively high resolutions, e.g., at resolutions better than the wavelength of visible light. The signaling entity may be, for example, a dye, a small molecule, a peptide or protein, or the like. The signaling entity may be a single molecule in some cases. If multiple secondary nucleic acid probes are used, the nucleic acid probes may comprise the same or different signaling entities.

Non-limiting examples of signaling entities include fluorescent entities (fluorophores) or phosphorescent entities, for example, cyanine dyes (e.g., Cy2, Cy3, Cy3B, Cy5, Cy5.5, Cy7, etc.), Alexa Fluor dyes, Atto dyes, photoswitchable dyes, photoactivatable dyes, fluorescent dyes, metal nanoparticles, semiconductor nanoparticles or “quantum dots”, fluorescent proteins such as GFP (Green Fluorescent Protein), or photoactivatable fluorescent proteins, such as PAGFP, PSCFP, PSCFP2, Dendra, Dendra2, EosFP, tdEos, mEos2, mEos3, PAmCherry, PAtagRFP, mMaple, mMaple2, and mMaple3. Other suitable signaling entities are known to those of ordinary skill in the art. See, e.g., U.S. Pat. No. 7,838,302 or U.S. Pat. Apl. Ser. No. 61/979,436, each incorporated herein by reference in its entirety.

As used herein, the term “light” generally refers to electromagnetic radiation, having any suitable wavelength (or equivalently, frequency). For instance, in some embodiments, the light may include wavelengths in the optical or visual range (for example, having a wavelength of between about 400 nm and about 700 nm, i.e., “visible light”), infrared wavelengths (for example, having a wavelength of between about 300 micrometers and 700 nm), ultraviolet wavelengths (for example, having a wavelength of between about 400 nm and about 10 nm), or the like. In certain cases, as discussed in detail below, more than one entity may be used, i.e., entities that are chemically different or distinct, for example, structurally. However, in other cases, the entities may be chemically identical or at least substantially chemically identical.

In one set of embodiments, the signaling entity is “switchable,” i.e., the entity can be switched between two or more states, at least one of which emits light having a desired wavelength. In the other state(s), the entity may emit no light, or emit light at a different wavelength. For instance, an entity may be “activated” to a first state able to produce light having a desired wavelength, and “deactivated” to a second state not able to emit light of the same wavelength. An entity is “photoactivatable” if it can be activated by incident light of a suitable wavelength. As a non-limiting example, Cy5, can be switched between a fluorescent and a dark state in a controlled and reversible manner by light of different wavelengths, i.e., 633

nm (or 642nm, 647nm, 656 nm) red light can switch or deactivate Cy5 to a stable dark state, while 405 nm green light can switch or activate the Cy5 back to the fluorescent state. In some cases, the entity can be reversibly switched between the two or more states, e.g., upon exposure to the proper stimuli. For example, a first stimuli (e.g., a first wavelength of light) may be used to activate the switchable entity, while a second stimuli (e.g., a second wavelength of light) may be used to deactivate the switchable entity, for instance, to a non-emitting state. Any suitable method may be used to activate the entity. For example, in one embodiment, incident light of a suitable wavelength may be used to activate the entity to emit light, i.e., the entity is “photoswitchable.” Thus, the photoswitchable entity can be switched between different light-emitting or non-emitting states by incident light, e.g., of different wavelengths. The light may be monochromatic (e.g., produced using a laser) or polychromatic. In another embodiment, the entity may be activated upon stimulation by electric field and/or magnetic field. In other embodiments, the entity may be activated upon exposure to a suitable chemical environment, e.g., by adjusting the pH, or inducing a reversible chemical reaction involving the entity, etc. Similarly, any suitable method may be used to deactivate the entity, and the methods of activating and deactivating the entity need not be the same. For instance, the entity may be deactivated upon exposure to incident light of a suitable wavelength, or the entity may be deactivated by waiting a sufficient time.

Typically, a “switchable” entity can be identified by one of ordinary skill in the art by determining conditions under which an entity in a first state can emit light when exposed to an excitation wavelength, switching the entity from the first state to the second state, e.g., upon exposure to light of a switching wavelength, then showing that the entity, while in the second state can no longer emit light (or emits light at a much reduced intensity) when exposed to the excitation wavelength.

In one set of embodiments, as discussed, a switchable entity may be switched upon exposure to light. In some cases, the light used to activate the switchable entity may come from an external source, e.g., a light source such as a laser light source, another light-emitting entity proximate the switchable entity, etc. The second, light emitting entity, in some cases, may be a fluorescent entity, and in certain embodiments, the second, light-emitting entity may itself also be a switchable entity.

In some embodiments, the switchable entity includes a first, light-emitting portion (e.g., a fluorophore), and a second portion that activates or “switches” the first portion. For example, upon exposure to light, the second portion of the switchable entity may activate the first portion, causing the first portion to emit light. Examples of activator portions include, but are not limited to, Alexa Fluor 405 (Invitrogen), Alexa Fluor 488 (Invitrogen), Cy2 (GE Healthcare), Cy3 (GE Healthcare), Cy3B (GE Healthcare), Cy3.5 (GE Healthcare), or other suitable dyes. Examples of light-emitting portions include, but are not limited to, Cy5, Cy5.5 (GE Healthcare), Cy7 (GE Healthcare), Alexa Fluor 647 (Invitrogen), Alexa Fluor 680 (Invitrogen), Alexa Fluor 700 (Invitrogen), Alexa Fluor 750 (Invitrogen), Alexa Fluor 790 (Invitrogen), DiD, DiR, YOYO-3 (Invitrogen), YO-PRO-3 (Invitrogen), TOT-3 (Invitrogen), TO-PRO-3 (Invitrogen) or other suitable dyes. These may be linked together, e.g., covalently, for example, directly, or through a linker, e.g., forming compounds such as, but not limited to, Cy5-Alexa Fluor 405, Cy5-Alexa Fluor 488, Cy5-Cy2, Cy5-Cy3, Cy5-Cy3.5, Cy5.5-Alexa Fluor 405, Cy5.5-Alexa Fluor 488, Cy5.5-Cy2, Cy5.5-Cy3, Cy5.5-Cy3.5, Cy7-Alexa Fluor 405, Cy7-Alexa Fluor 488, Cy7-Cy2, Cy7-Cy3, Cy7-Cy3.5, Alexa Fluor 647-Alexa Fluor 405, Alexa Fluor 647-Alexa Fluor 488, Alexa Fluor 647-Cy2, Alexa Fluor 647-Cy3, Alexa Fluor 647-Cy3.5, Alexa Fluor 750-Alexa Fluor 405, Alexa Fluor 750-Alexa Fluor 488, Alexa Fluor 750-Cy2, Alexa Fluor 750-Cy3, or Alexa Fluor 750-Cy3.5. . Those of ordinary skill in the art will be aware of the structures of these and other compounds, many of which are available commercially. The portions may be linked via a covalent bond, or by a linker, such as those described in detail below. Other light-emitting or activator portions may include portions having two quaternized nitrogen atoms joined by a polymethine chain, where each nitrogen is independently part of a heteroaromatic moiety, such as pyrrole, imidazole, thiazole, pyridine, quinoine, indole, benzothiazole, etc., or part of a nonaromatic amine. In some cases, there may be 5, 6, 7, 8, 9, or more carbon atoms between the two nitrogen atoms.

In certain cases, the light-emitting portion and the activator portions, when isolated from each other, may each be fluorophores, i.e., entities that can emit light of a certain, emission wavelength when exposed to a stimulus, for example, an excitation wavelength. However, when a switchable entity is formed that comprises the first fluorophore and the second fluorophore, the first fluorophore forms a first, light-emitting portion and the second fluorophore forms an activator portion that switches that activates or “switches” the first portion in response to a

stimulus. For example, the switchable entity may comprise a first fluorophore directly bonded to the second fluorophore, or the first and second entity may be connected via a linker or a common entity. Whether a pair of light-emitting portion and activator portion produces a suitable switchable entity can be tested by methods known to those of ordinary skills in the art. For example, light of various wavelength can be used to stimulate the pair and emission light from the light-emitting portion can be measured to determine whether the pair makes a suitable switch.

As a non-limiting example, Cy3 and Cy5 may be linked together to form such an entity. In this example, Cy3 is an activator portion that is able to activate Cy5, the light-emission portion. Thus, light at or near the absorption maximum (e.g., near 532 nm light for Cy3) of the activation or second portion of the entity may cause that portion to activate the first, light-emitting portion, thereby causing the first portion to emit light (e.g., near 647 nm for Cy5). See, e.g., U.S. Pat. No. 7,838,302, incorporated herein by reference in its entirety. In some cases, the first, light-emitting portion can subsequently be deactivated by any suitable technique (e.g., by directing 647 nm red light to the Cy5 portion of the molecule).

Other non-limiting examples of potentially suitable activator portions include 1,5-IAEDANS, 1,8-ANS, 4-Methylumbelliferone, 5-carboxy-2,7-dichlorofluorescein, 5-Carboxyfluorescein (5-FAM), 5-Carboxynaphthofluorescein, 5-Carboxytetramethylrhodamine (5-TAMRA), 5-FAM (5-Carboxyfluorescein), 5-HAT (Hydroxy Tryptamine), 5-Hydroxy Tryptamine (HAT), 5-ROX (carboxy-X-rhodamine), 5-TAMRA (5-Carboxytetramethylrhodamine), 6-Carboxyrhodamine 6G, 6-CR 6G, 6-JOE, 7-Amino-4-methylcoumarin, 7-Aminoactinomycin D (7-AAD), 7-Hydroxy-4-methylcoumarin, 9-Amino-6-chloro-2-methoxyacridine, ABQ, Acid Fuchsin, ACMA (9-Amino-6-chloro-2-methoxyacridine), Acridine Orange, Acridine Red, Acridine Yellow, Acriflavin, Acriflavin Feulgen SITSA, Alexa Fluor 350, Alexa Fluor 405, Alexa Fluor 430, Alexa Fluor 488, Alexa Fluor 500, Alexa Fluor 514, Alexa Fluor 532, Alexa Fluor 546, Alexa Fluor 555, Alexa Fluor 568, Alexa Fluor 594, Alexa Fluor 610, Alexa Fluor 633, Alexa Fluor 635, Alizarin Complexon, Alizarin Red, AMC, AMCA-S, AMCA (Aminomethylcoumarin), AMCA-X, Aminoactinomycin D, Aminocoumarin, Aminomethylcoumarin (AMCA), Anilin Blue, Anthrocyl stearate, APTRA-BTC, APTS, Astrazon Brilliant Red 4G, Astrazon Orange R, Astrazon Red 6B, Astrazon Yellow 7 GLL, Atabrine, ATTO 390, ATTO 425, ATTO 465, ATTO 488, ATTO 495, ATTO 520, ATTO 532, ATTO 550, ATTO 565, ATTO 590, ATTO 594, ATTO 610, ATTO 611X, ATTO 620, ATTO

633, ATTO 635, ATTO 647, ATTO 647N, ATTO 655, ATTO 680, ATTO 700, ATTO 725, ATTO 740, ATTO-TAG CBQCA, ATTO-TAG FQ, Auramine, Aurophosphine G, Aurophosphine, BAO 9 (Bisaminophenyloxadiazole), BCECF (high pH), BCECF (low pH), Berberine Sulphate, Bimane, Bisbenzamide, Bisbenzimidazole (Hoechst), bis-BTC, Blancophor

5 FFG, Blancophor SV, BOBO -1, BOBO -3, Bodipy 492/515, Bodipy 493/503, Bodipy 500/510, Bodipy 505/515, Bodipy 530/550, Bodipy 542/563, Bodipy 558/568, Bodipy 564/570, Bodipy 576/589, Bodipy 581/591, Bodipy 630/650-X, Bodipy 650/665-X, Bodipy 665/676, Bodipy FL, Bodipy FL ATP, Bodipy FL-Ceramide, Bodipy R6G, Bodipy TMR, Bodipy TMR-X conjugate, Bodipy TMR-X, SE, Bodipy TR, Bodipy TR ATP, Bodipy TR-X SE, BO-PRO -1, BO-PRO -3,

10 Brilliant Sulphoflavin FF, BTC, BTC-5N, Calcein, Calcein Blue, Calcium Crimson, Calcium Green, Calcium Green-1 Ca^{2+} Dye, Calcium Green-2 Ca^{2+} , Calcium Green-5N Ca^{2+} , Calcium Green-C18 Ca^{2+} , Calcium Orange, Calcofluor White, Carboxy-X-rhodamine (5-ROX), Cascade Blue, Cascade Yellow, Catecholamine, CCF2 (GeneBlazer), CFDA, Chromomycin A, Chromomycin A, CL-NERF, CMFDA, Coumarin Phalloidin, CPM Methylcoumarin, CTC, CTC

15 Formazan, Cy2, Cy3.1 8, Cy3.5, Cy3, Cy5.1 8, cyclic AMP Fluorosensor (FiCRhR), Dabcyl, Dansyl, Dansyl Amine, Dansyl Cadaverine, Dansyl Chloride, Dansyl DHPE, Dansyl fluoride, DAPI, Dapoxyl, Dapoxyl 2, Dapoxyl 3' DCFDA, DCFH (Dichlorodihydrofluorescein Diacetate), DDAO, DHR (Dihydorhodamine 123), Di-4-ANEPPS, Di-8-ANEPPS (non-ratio), DiA (4-Di-16-ASP), Dichlorodihydrofluorescein Diacetate (DCFH), DiD - Lipophilic Tracer, DiD

20 (DiIC18(5)), DIDS, Dihydorhodamine 123 (DHR), DiI (DiIC18(3)), Dinitrophenol, DiO (DiOC18(3)), DiR, DiR (DiIC18(7)), DM-NERF (high pH), DNP, Dopamine, DTAF, DY-630-NHS, DY-635-NHS, DyLight 405, DyLight 488, DyLight 549, DyLight 633, DyLight 649, DyLight 680, DyLight 800, ELF 97, Eosin, Erythrosin, Erythrosin ITC, Ethidium Bromide, Ethidium homodimer -1 (EthD-1), Euchrysin, EukoLight, Europium (III) chloride, Fast Blue,

25 FDA, Feulgen (Pararosaniline), FIF (Formaldehyd Induced Fluorescence), FITC, Flazo Orange, Fluo-3, Fluo-4, Fluorescein (FITC), Fluorescein Diacetate, Fluoro-Emerald, Fluoro-Gold (Hydroxystilbamidine), Fluor-Ruby, FluorX, FM 1-43, FM 4-46, Fura Red (high pH), Fura Red/Fluo-3, Fura-2, Fura-2/BCECF, Genacryl Brilliant Red B, Genacryl Brilliant Yellow 10GF, Genacryl Pink 3G, Genacryl Yellow 5GF, GeneBlazer (CCF2), Gloxalic Acid, Granular blue,

30 Haematoporphyrin, Hoechst 33258, Hoechst 33342, Hoechst 34580, HPTS, Hydroxycoumarin, Hydroxystilbamidine (FluoroGold), Hydroxytryptamine, Indo-1, high calcium, Indo-1, low

calcium, Indodicarbocyanine (DiD), Indotricarbocyanine (DiR), Intrawhite Cf, JC-1, JO-JO-1, JO-PRO-1, LaserPro, Laurodan, LDS 751 (DNA), LDS 751 (RNA), Leucophor PAF, Leucophor SF, Leucophor WS, Lissamine Rhodamine, Lissamine Rhodamine B, Calcein/Ethidium homodimer, LOLO-1, LO-PRO-1, Lucifer Yellow, Lyso Tracker Blue, Lyso Tracker Blue-
5 White, Lyso Tracker Green, Lyso Tracker Red, Lyso Tracker Yellow, LysoSensor Blue, LysoSensor Green, LysoSensor Yellow/Blue, Mag Green, Magdala Red (Phloxin B), Mag-Fura Red, Mag-Fura-2, Mag-Fura-5, Mag-Indo-1, Magnesium Green, Magnesium Orange, Malachite Green, Marina Blue, Maxilon Brilliant Flavin 10 GFF, Maxilon Brilliant Flavin 8 GFF, Merocyanin, Methoxycoumarin, Mitotracker Green FM, Mitotracker Orange, Mitotracker Red,
10 Mitramycin, Monobromobimane, Monobromobimane (mBBr-GSH), Monochlorobimane, MPS (Methyl Green Pyronine Stilbene), NBD, NBD Amine, Nile Red, Nitrobenzoxadidole, Noradrenaline, Nuclear Fast Red, Nuclear Yellow, Nylosan Brilliant Iavin E8G, Oregon Green, Oregon Green 488-X, Oregon Green, Oregon Green 488, Oregon Green 500, Oregon Green 514, Pacific Blue, Pararosanine (Feulgen), PBFI, Phloxin B (Magdala Red), Phorwite AR, Phorwite
15 BKL, Phorwite Rev, Phorwite RPA, Phosphine 3R, PKH26 (Sigma), PKH67, PMIA, Pontochrome Blue Black, POPO-1, POPO-3, PO-PRO-1, PO-PRO-3, Primuline, Procion Yellow, Propidium Iodid (PI), PyMPO, Pyrene, Pyronine, Pyronine B, Pyroزال Brilliant Flavin 7GF, QSY 7, Quinacrine Mustard, Resorufin, RH 414, Rhod-2, Rhodamine, Rhodamine 110, Rhodamine 123, Rhodamine 5 GLD, Rhodamine 6G, Rhodamine B, Rhodamine B 200,
20 Rhodamine B extra, Rhodamine BB, Rhodamine BG, Rhodamine Green, Rhodamine Phallicidine, Rhodamine Phalloidine, Rhodamine Red, Rhodamine WT, Rose Bengal, S65A, S65C, S65L, S65T, SBFI, Serotonin, Sevron Brilliant Red 2B, Sevron Brilliant Red 4G, Sevron Brilliant Red B, Sevron Orange, Sevron Yellow L, SITS, SITS (Primuline), SITS (Stilbene Isothiosulphonic Acid), SNAFL calcein, SNAFL-1, SNAFL-2, SNARF calcein, SNARF1,
25 Sodium Green, SpectrumAqua, SpectrumGreen, SpectrumOrange, Spectrum Red, SPQ (6-methoxy-N-(3-sulfopropyl)quinolinium), Stilbene, Sulphorhodamine B can C, Sulphorhodamine Extra, SYTO 11, SYTO 12, SYTO 13, SYTO 14, SYTO 15, SYTO 16, SYTO 17, SYTO 18, SYTO 20, SYTO 21, SYTO 22, SYTO 23, SYTO 24, SYTO 25, SYTO 40, SYTO 41, SYTO 42, SYTO 43, SYTO 44, SYTO 45, SYTO 59, SYTO 60, SYTO 61, SYTO 62, SYTO 63,
30 SYTO 64, SYTO 80, SYTO 81, SYTO 82, SYTO 83, SYTO 84, SYTO 85, SYTOX Blue, SYTOX Green, SYTOX Orange, Tetracycline, Tetramethylrhodamine (TAMRA), Texas Red,

Texas Red-X conjugate, Thiadiazine Red R, Thiazole Orange, Thioflavin 5, Thioflavin S, Thioflavin TCN, Thiolyte, Thiozole Orange, Tinopol CBS (Calcofluor White), TMR, TO-PRO-1, TO-PRO-3, TO-PRO-5, TOTO-1, TOTO-3, TRITC (tetramethylrhodamine isothiocyanate), True Blue, TruRed, Ultralite, Uranine B, Uvitex SFC, WW 781, X-Rhodamine, XRITC, Xylene Orange, Y66F, Y66H, Y66W, YO-PRO-1, YO-PRO-3, YOYO-1, YOYO-3, SYBR Green, Thiazole orange (interchelating dyes), or combinations thereof.

Another aspect of the invention is directed to a computer-implemented method. For instance, a computer and/or an automated system may be provided that is able to automatically and/or repetitively perform any of the methods described herein. As used herein, “automated” devices refer to devices that are able to operate without human direction, i.e., an automated device can perform a function during a period of time after any human has finished taking any action to promote the function, e.g. by entering instructions into a computer to start the process. Typically, automated equipment can perform repetitive functions after this point in time. The processing steps may also be recorded onto a machine-readable medium in some cases.

For example, in some cases, a computer may be used to control imaging of the sample, e.g., using STORM or other super-resolution techniques such as those described herein. In some cases, the computer may also control operations such as drift correction, physical registration, hybridization and cluster alignment in image analysis, cluster decoding, error detection or correction (e.g., as discussed herein), noise reduction, identification of foreground features from background features (such as noise or debris in images), or the like. As an example, the computer may be used to control activation and/or excitation of signaling entities within the sample, and/or the acquisition of images of the signaling entities. In one set of embodiments, a sample may be excited using light having various wavelengths and/or intensities, and the sequence of the wavelengths of light used to excite the sample may be correlated, using a computer, to the images acquired of the sample containing the signaling entities. For instance, the computer may apply light having various wavelengths and/or intensities to a sample to yield different average numbers of signaling entities in each region of interest (e.g., one activated entity per location, two activated entities per location, etc.). In some cases, this information may be used to construct an image and/or determine the locations of the signaling entities, in some cases at high resolutions, as noted above.

In some aspects, the sample is positioned on a microscope. In some cases, the microscope may contain one or more channels, such as microfluidic channels, to direct or control fluid to or from the sample. For instance, in one embodiment, nucleic acid probes such as those discussed herein may be introduced and/or removed from the sample by flowing fluid through one or more channels to or from the sample. In some cases, there may also be one or more chambers or reservoirs for holding fluid, e.g., in fluidic communication with the channel, and/or with the sample. Those of ordinary skill in the art will be familiar with channels, including microfluidic channels, for moving fluid to or from a sample.

As used herein, “microfluidic,” “microscopic,” “microscale,” the “micro-” prefix (for example, as in “microchannel”), and the like generally refers to elements or articles having widths or diameters of less than about 1 mm, and less than about 100 microns (micrometers) in some cases. In some embodiments, larger channels may be used instead of, or in conjunction with, microfluidic channels for any of the embodiments discussed herein. For examples, channels having widths or diameters of less than about 10 mm, less than about 9 mm, less than about 8 mm, less than about 7 mm, less than about 6 mm, less than about 5 mm, less than about 4 mm, less than about 3 mm, or less than about 2 mm may be used in certain instances. In some cases, the element or article includes a channel through which a fluid can flow. In all embodiments, specified widths can be a smallest width (i.e. a width as specified where, at that location, the article can have a larger width in a different dimension), or a largest width (i.e. where, at that location, the article has a width that is no wider than as specified, but can have a length that is greater). Thus, for instance, the microfluidic channel may have an average cross-sectional dimension (e.g., perpendicular to the direction of flow of fluid in the microfluidic channel) of less than about 1 mm, less than about 500 microns, less than about 300 microns, or less than about 100 microns. In some cases, the microfluidic channel may have an average diameter of less than about 60 microns, less than about 50 microns, less than about 40 microns, less than about 30 microns, less than about 25 microns, less than about 10 microns, less than about 5 microns, less than about 3 microns, or less than about 1 micron.

A “channel,” as used herein, means a feature on or in an article (e.g., a substrate) that at least partially directs the flow of a fluid. In some cases, the channel may be formed, at least in part, by a single component, e.g. an etched substrate or molded unit. The channel can have any cross-sectional shape, for example, circular, oval, triangular, irregular, square or rectangular

(having any aspect ratio), or the like, and can be covered or uncovered (i.e., open to the external environment surrounding the channel). In embodiments where the channel is completely covered, at least one portion of the channel can have a cross-section that is completely enclosed, and/or the entire channel may be completely enclosed along its entire length with the exception of its inlet and outlet.

A channel may have any aspect ratio, e.g., an aspect ratio (length to average cross-sectional dimension) of at least about 2:1, more typically at least about 3:1, at least about 5:1, at least about 10:1, etc. As used herein, a “cross-sectional dimension,” in reference to a fluidic or microfluidic channel, is measured in a direction generally perpendicular to fluid flow within the channel. A channel generally will include characteristics that facilitate control over fluid transport, e.g., structural characteristics and/or physical or chemical characteristics (hydrophobicity vs. hydrophilicity) and/or other characteristics that can exert a force (e.g., a containing force) on a fluid. The fluid within the channel may partially or completely fill the channel. In some cases the fluid may be held or confined within the channel or a portion of the channel in some fashion, for example, using surface tension (e.g., such that the fluid is held within the channel within a meniscus, such as a concave or convex meniscus). In an article or substrate, some (or all) of the channels may be of a particular size or less, for example, having a largest dimension perpendicular to fluid flow of less than about 5 mm, less than about 2 mm, less than about 1 mm, less than about 500 microns, less than about 200 microns, less than about 100 microns, less than about 60 microns, less than about 50 microns, less than about 40 microns, less than about 30 microns, less than about 25 microns, less than about 10 microns, less than about 3 microns, less than about 1 micron, less than about 300 nm, less than about 100 nm, less than about 30 nm, or less than about 10 nm or less in some cases. In one embodiment, the channel is a capillary.

A variety of materials and methods, according to certain aspects of the invention, can be used to form devices or components containing microfluidic channels, chambers, etc. For example, various devices or components can be formed from solid materials, in which the channels can be formed via micromachining, film deposition processes such as spin coating and chemical vapor deposition, physical vapor deposition, laser fabrication, photolithographic techniques, etching methods including wet chemical or plasma processes, electrodeposition, and the like. See, for example, *Scientific American*, 248:44-55, 1983 (Angell, *et al*).

In one set of embodiments, various structures or components can be formed of a polymer, for example, an elastomeric polymer such as polydimethylsiloxane ("PDMS"), polytetrafluoroethylene ("PTFE" or Teflon[®]), or the like. For instance, according to one embodiment, a channel such as a microfluidic channel may be implemented by fabricating the fluidic system separately using PDMS or other soft lithography techniques (details of soft lithography techniques suitable for this embodiment are discussed in the references entitled "Soft Lithography," by Younan Xia and George M. Whitesides, published in the *Annual Review of Material Science*, 1998, Vol. 28, pages 153-184, and "Soft Lithography in Biology and Biochemistry," by George M. Whitesides, Emanuele Ostuni, Shuichi Takayama, Xingyu Jiang and Donald E. Ingber, published in the *Annual Review of Biomedical Engineering*, 2001, Vol. 3, pages 335-373; each of these references is incorporated herein by reference).

Other examples of potentially suitable polymers include, but are not limited to, polyethylene terephthalate (PET), polyacrylate, polymethacrylate, polycarbonate, polystyrene, polyethylene, polypropylene, polyvinylchloride, cyclic olefin copolymer (COC), polytetrafluoroethylene, a fluorinated polymer, a silicone such as polydimethylsiloxane, polyvinylidene chloride, bis-benzocyclobutene ("BCB"), a polyimide, a fluorinated derivative of a polyimide, or the like. Combinations, copolymers, or blends involving polymers including those described above are also envisioned. The device may also be formed from composite materials, for example, a composite of a polymer and a semiconductor material.

In some embodiments, various microfluidic structures or components of the device are fabricated from polymeric and/or flexible and/or elastomeric materials, and can be conveniently formed of a hardenable fluid, facilitating fabrication via molding (e.g. replica molding, injection molding, cast molding, etc.). The hardenable fluid can be essentially any fluid that can be induced to solidify, or that spontaneously solidifies, into a solid capable of containing and/or transporting fluids contemplated for use in and with the fluidic network. In one embodiment, the hardenable fluid comprises a polymeric liquid or a liquid polymeric precursor (i.e. a "prepolymer"). Suitable polymeric liquids can include, for example, thermoplastic polymers, thermoset polymers, waxes, metals, or mixtures or composites thereof heated above their melting point. As another example, a suitable polymeric liquid may include a solution of one or more polymers in a suitable solvent, which solution forms a solid polymeric material upon removal of the solvent, for example, by evaporation. Such polymeric materials, which can be solidified

from, for example, a melt state or by solvent evaporation, are well known to those of ordinary skill in the art. A variety of polymeric materials, many of which are elastomeric, are suitable, and are also suitable for forming molds or mold masters, for embodiments where one or both of the mold masters is composed of an elastomeric material. A non-limiting list of examples of such polymers includes polymers of the general classes of silicone polymers, epoxy polymers, and acrylate polymers. Epoxy polymers are characterized by the presence of a three-membered cyclic ether group commonly referred to as an epoxy group, 1,2-epoxide, or oxirane. For example, diglycidyl ethers of bisphenol A can be used, in addition to compounds based on aromatic amine, triazine, and cycloaliphatic backbones. Another example includes the well-known Novolac polymers. Non-limiting examples of silicone elastomers suitable for use according to the invention include those formed from precursors including the chlorosilanes such as methylchlorosilanes, ethylchlorosilanes, phenylchlorosilanes, etc.

Silicone polymers are used in certain embodiments, for example, the silicone elastomer polydimethylsiloxane. Non-limiting examples of PDMS polymers include those sold under the trademark Sylgard by Dow Chemical Co., Midland, MI, and particularly Sylgard 182, Sylgard 184, and Sylgard 186. Silicone polymers including PDMS have several beneficial properties simplifying fabrication of various structures of the invention. For instance, such materials are inexpensive, readily available, and can be solidified from a prepolymeric liquid via curing with heat. For example, PDMSs are typically curable by exposure of the prepolymeric liquid to temperatures of about, for example, about 65 °C to about 75 °C for exposure times of, for example, at least about an hour. Also, silicone polymers, such as PDMS, can be elastomeric and thus may be useful for forming very small features with relatively high aspect ratios, necessary in certain embodiments of the invention. Flexible (e.g., elastomeric) molds or masters can be advantageous in this regard.

One advantage of forming structures such as microfluidic structures or channels from silicone polymers, such as PDMS, is the ability of such polymers to be oxidized, for example by exposure to an oxygen-containing plasma such as an air plasma, so that the oxidized structures contain, at their surface, chemical groups capable of cross-linking to other oxidized silicone polymer surfaces or to the oxidized surfaces of a variety of other polymeric and non-polymeric materials. Thus, structures can be fabricated and then oxidized and essentially irreversibly sealed to other silicone polymer surfaces, or to the surfaces of other substrates reactive with the

oxidized silicone polymer surfaces, without the need for separate adhesives or other sealing means. In most cases, sealing can be completed simply by contacting an oxidized silicone surface to another surface without the need to apply auxiliary pressure to form the seal. That is, the pre-oxidized silicone surface acts as a contact adhesive against suitable mating surfaces.

5 Specifically, in addition to being irreversibly sealable to itself, oxidized silicone such as oxidized PDMS can also be sealed irreversibly to a range of oxidized materials other than itself including, for example, glass, silicon, silicon oxide, quartz, silicon nitride, polyethylene, polystyrene, glassy carbon, and epoxy polymers, which have been oxidized in a similar fashion to the PDMS surface (for example, via exposure to an oxygen-containing plasma). Oxidation and sealing methods
10 useful in the context of the present invention, as well as overall molding techniques, are described in the art, for example, in an article entitled “Rapid Prototyping of Microfluidic Systems and Polydimethylsiloxane,” *Anal. Chem.*, 70:474-480, 1998 (Duffy *et al.*), incorporated herein by reference.

The following documents are each incorporated herein by reference in their entireties:

15 U.S. Pat. No. 7,838,302, issued November 23, 2010, entitled “Sub-Diffraction Limit Image Resolution and Other Imaging Techniques,” by Zhuang, *et al.*; U.S. Pat. No. 8,564,792, issued October 22, 2013, entitled “Sub-diffraction Limit Image Resolution in Three Dimensions,” by Zhuang, *et al.*; and Int. Pat. Apl. Pub. No. WO 2013/090360, published June 20, 2013, entitled “High Resolution Dual-Objective Microscopy,” by Zhuang, *et al.*

20 The following examples are intended to illustrate certain embodiments of the present invention, but do not exemplify the full scope of the invention.

EXAMPLE 1

The example presents a platform to enable the simultaneous detection of the number and spatial organization of thousands of distinct mRNAs within single cells with high efficiency and
25 low error-rate using a novel form of highly multiplexed fluorescence *in situ* hybridization (FISH). This example accomplishes these measurements by integrating and innovating methods for massively parallel probe synthesis, super-resolution imaging, and self-correcting error-checking codes.

Here, these examples present methods for the simultaneous detection of some or all of the
30 thousands of unique RNAs expressed in a cell. This approach not only promises to revolutionize the throughput of the already effective smFISH approach, but also allows researchers to benefit

from the hypothesis free discovery approach which has made other whole-genome systems approaches to biology so effective. For example, this whole genome approach may allow researchers to discover RNAs whose expression levels and/or subcellular localization patterns change under certain conditions of interest, such as disease states, without knowing, *a priori*, which mRNA will change in abundance or localization. Simultaneous measurements of hundreds of genes within a single cell also allow for the identification of correlations between genes in expression and localization patterns in some cases.

This can be achieved using methods for highly multiplexed smFISH via the sequential hybridization of orthogonal detection probes and super-resolution imaging, reducing the cost of probe synthesis, and the development of a highly automated system to minimize demands on the user, as discussed herein. This provides an integrated platform to handle the bioinformatics of probe design, the mathematics of error-correcting codes, the complexity of image registration and analysis, and the cumbersome fluid handling through a simple suite of user friendly interfaces. This integration allows easy operation with limited user training and facilitates the rapid collection of data.

This example illustrates: (1) computational design of “codewords” to attach to all RNA targets in the cell that will allow unique identification of each RNA with some degree of experimental error tolerance, (2) translation of these codewords into nucleotide sequences and synthesis of the required single-stranded (ss) oligonucleotide (e.g. ssDNA) probes, (3) sample fixation and hybridization of these probes to the RNA targets *in situ*, (4) read-out of these codewords via successive rounds of hybridization of distinct fluorescent probes imaged with conventional fluorescence microscopy or super-resolution fluorescence microscopy, and (5) automated decoding of measured codewords combined with computational error correction to uniquely and robustly identify individual mRNAs.

In the first step, a “codeword” is assigned to every RNA that is to be labeled. In a typical design these may be strings of N binary letters or positions. Codewords may be chosen from the same wide range of existing error tolerant or error-correcting encoding schemes developed for digital storage and communication, e.g., using Hamming codes or the like. For example, actin-RNA may be assigned the binary codeword 11001010. Each codeword may be unique and separated from the other codewords by a Hamming distance, h , which measures the number of letters or positions that must be incorrectly read for one codeword to be misinterpreted as a

different one. A Hamming distance greater than 1 between all codewords allows for some measurement errors to be detected—since simple errors would produce codewords that are not used to encode RNAs. For a Hamming distance larger than 2, it is also possible to correct some errors, as codewords with one error will be closest in Hamming distance to a single, unique
5 codeword. The total number of different RNAs to be detected from the transcriptome and the amount of error correction desired determines the length of the codewords. Information theory provides several efficient algorithms for assembling error-correcting binary codebooks.

In the second step, this encoding scheme is translated into a set of oligonucleotide (e.g. DNA) probe sequences, each of which not only targets a probe to the RNA of interest but also
10 encodes the unique binary codeword within a set of secondary binding sites (Fig. 1). For example, first designed may be primary binding sequences for each targeted mRNA. These sequences are “target sequences” that are comprised of complementary nucleotide sequences to their target RNAs computationally selected to satisfy a stringent set of hybridization conditions, including uniqueness in the target genome. To improve the efficiency of hybridization to
15 individual mRNAs, multiple primary target sequences are designed for each individual RNA. Then each position within the set of codewords is assigned a unique oligonucleotide (e.g. DNA) sequence, which is called a read sequence. These tags are designed as to have no interaction with endogenous mRNA sequences or each other. For instance, for all the value “1”s in a codeword of an individual mRNA, the corresponding read sequence is attached to the primary
20 targeting sequences against that mRNA. In general, each probe will contain a target sequence and one or more read sequences. If the total length of the necessary read sequences and the primary target sequence exceeds synthesis capabilities, then subsets of the read sequences can be appended to distinct target sequences. For example, consider the potential codeword 11001010 for actin. Probe sequences for this RNA could contain the read sequences corresponding to
25 positions 1, 2, 5, and 7 in the codeword attached to a variety of target sequences specific to actin. After all the sequences have been designed, the resulting complex set of unique custom oligonucleotide (e.g. DNA) sequences is manufactured and amplified using methods as described below.

In the third step, the resulting pool of DNA is hybridized, e.g., to fixed, permeabilized
30 cells. In this process, individual probes may be attached to every RNA in the cell by

hybridization of their corresponding target sequences with the RNA while the read sequences remain free to bind the appropriate secondary probes as discussed below.

In the fourth step—the read-out step—fluorescently labeled secondary nucleic acid probes are successively hybridized to the read sequences attached to the target sequences that
5 binds to the mRNA targets in the above step. Even for a modest number of moderately expressed mRNAs, the density of labeled mRNAs exceed that at which each mRNA can be resolved via conventional imaging methods. Thus, this may be performed using a super-resolution imaging method, for example STORM (stochastic optical reconstruction microscopy), to resolve the labeled molecules. After each round of secondary hybridization and imaging, the
10 fluorophores are quenched or otherwise inactivated either via chemical or optical techniques such as oxidation, chemically bleaching, photobleaching, stringent washing or enzymatic digestion, etc. The sample is then stained with the next secondary probe, and the cycle continues until all positions of the codewords have been read out. In the simplest incarnation, there will be one hybridization step for each position within the codeword, e.g. 8 hybridization steps for an 8-
15 letter codeword (Fig. 1).

Fig. 1 shows schematic diagrams of this example. Fig. 1A shows that every position of the codewords is assigned a unique oligonucleotide sequence when this position has a value “1.” All mRNA codewords are then translated into combinations of read sequences, which are attached to the targeting sequence. Fig. 1B shows various steps of the labeling scheme of this
20 example. In the first step, all mRNAs (I - III) are tagged with multiple oligonucleotide (e.g. ssDNA) probes comprising a primary targeting sequence, which hybridizes to the RNA of interest, and a “tail” (i.e. containing read sequences) carrying the translated codeword, which does not interact with endogenous nucleotide sequences. In the next step, the first secondary probe is added, which can bind all probes whose tails have a read sequence corresponding to the
25 value of “1” in the first position. The dyes on these secondary probes are imaged and bleached, then the next secondary probe is added to bind probes attached to mRNA which have a value of “1” in the second position of their assigned codeword, and so on.

In the final step, the microscopy images from each staining and imaging round are aligned, for example, computationally (e.g. using fiducial beads or other markers tracked during
30 image acquisition), and the clusters of localizations resolved by super-resolution imaging (e.g. STORM) from the different rounds are identified. These clusters of localizations arise from

individual target mRNA molecules, and the hybridization rounds in which a spot was detected in a given cluster correspond to the “1” in the codeword for that mRNA. If there are no missed-detection events or false positive signal in the images, this codeword will perfectly match one of the expected codewords. Fig. 1 describes an example in which the codeword has three letters, i.e. three positions, and the three target mRNAs have codewords 110, 101, and 011 assigned to them. In real experimental examples, the codeword could contain more digits. For example, the mRNA for actin can be assigned the codeword, 11001010. In that case, detected clusters containing overlapping localization signals in the 1st, 2nd, 5th and 7th hybridization steps (meaning the 1st, 2nd, 5th and 7th secondary probes bound to this site) can be identified as individual actin mRNA molecules, since the pattern of positive bindings matches the codeword of actin (11001010). In addition, if there are missed-detection events or false positive signals in the image data, these aberrations can be corrected by the implemented error-correction scheme. For example, clusters of localizations with a detected codeword that has only one digit discrepancy from 11001010 (such as 11000010 or 11101010) can also be identified as actin mRNA since all other valid codeword in this example differ from the detected pattern in two more positions.

EXAMPLE 2

This example describes another alternative approach that differs in several of the steps described above. This approach begins with the first step, construction of the codewords to the desired mRNA targets, as described above.

In the second step of this approach, nucleic acid probes are designed that bind uniquely to the mRNA targets of interest, as described above. However, instead of appending unique read sequences to these targeting sequences, unique pools or groups of probes are constructed from these target sequences. Each pool comprises all or a subset of the sequences that target all mRNAs which contain the same value at a given position in their codeword. For example, the first pool would have of all or a subset of the target sequences designed for all mRNAs that contain a 1 in the first position of their codewords, e.g. 110 and 101 but not 011; the second pool would have of all or a subset of the target sequences designed for all mRNAs that contain a 1 in the second position of their codewords, e.g. 110 and 011 but not 101; the third pool would have of all or a subset of the target sequences designed for all mRNAs that contain a 1 in the third position of their codewords, e.g. 011 and 101 but not 110 (Fig. 1C). As another example, consider the potential codeword 11001010 for actin. Probes that target this mRNA would be

included in pools 1, 2, 5, and 7 but not in pools 3, 4, 6, and 8. The same target to a given mRNA may or may not be included in pools. For example, a probe that targets the same region of actin may be included in pools 1, 2, 5, and 7 or any subset of these pools. After all of the pools have been designed, each complex set of unique, custom oligonucleotide sequences is manufactured and amplified using methods as described below.

In the third step of this approach, the first pool of probes is hybridized, e.g., to fixed, permeabilized cells. In this process, the fluorophores attached to each of the probes in this pool are bound to each of the targets of that pool. The binding of these probes is then determined by fluorescence microscopy. As described above, these images can be collected either via a range of methods including both conventional fluorescence imaging of super-resolution imaging methods such as STORM. After a round of imaging, the probes from the first pool are inactivated or removed from the sample via the methods described above. This process is then repeated for each successive pool of probes until some or all of the pools have been applied to the sample and imaged such that all positions in the codewords have been read out. In the simplest incarnation, there will be one hybridization and imaging step for each position in the codeword, e.g. 3 rounds of hybridization and imaging for a codeword with 3 positions (Fig 1C) or 8 rounds of hybridization and imaging for a codeword with 8 positions.

The final step of this approach is identical to that described above.

EXAMPLE 3

In this example, 14 genes (PGK1, H3F3B, PKM, ENO1, GPI, EEF2, GNAS, HSPA8, GAPDH, CALM1, RHOA, PPIA, UBA52, and VCP) were encoded using a subset of the (8,4) SECDED code (Figs. 2A-2E). To determine the accuracy of these measurements, the measured abundances of these 14 mRNAs were compared to the abundances measured from bulk RNA-seq of A549 cells (published ENCODE data). Remarkably, it was found that there was excellent agreement between these two measurements, as the transcripts count measured using the sequential hybridization approach correlated with gene expression measured using RNA-seq with a Pearson r of 0.75 (Fig. 2F). Gene expression from 3 other cells were also measured, and it was found that the gene expression of these 14 genes was highly correlated among the cells with a r of 0.96 (Fig. 2G).

Codebook Design. Each mRNA in the target set was assigned a binary codeword using a Single Error Correction Double Error Detection (SECDED) code. SECDED is an extended

Hamming codebook with an additional parity bit. Briefly, Matlab's Communications System toolbox was used to generate SECDED codes of either 8 or 16 letters or positions. In both cases, only those codewords containing four 1s were used. These words were assigned at random to mRNAs in the target set. [0 1 0 1 1 1 0 0] is an example of the 8-letter codewords used (i.e., these codewords each contained four 1s and four 0s.) [0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0] is an example of the 16-letter codewords used (i.e., each codeword contained four 1s and twelve 0s). Not every codeword was necessarily assigned to an mRNA.

Computational Assembly of ssDNA Primary Probe Sequence. The number of primary nucleic acid probes used for hybridization with mRNA targets ranged from 200 to 2000 unique oligonucleotides, depending on the experiment. For example, to label 14 mRNAs with 28 oligos targeting each gene, 392 unique sequences were used. Large number of oligos with unique sequences were purchased in a pool from LC Sciences or CustomArray. However, array synthesized oligos were in minute quantity that was insufficient for *in situ* hybridization. The protocol for their amplification are described below.

Each primary probe contained three components: flanking primer sequences to allow enzymatic amplification of probes, targeting sequence for *in situ* hybridization to mRNAs, and secondary tag sequence containing one or more read sequences for sequential readout of codewords.

The following is an example of a primary probe:

GTTGGCGACGAAAGCACTGCGATTGGAACCGTCCCAAGCGTTGCG
CTTAATGGATCATCAATTTTGTCTCACTACGACGGTCAATCGCGCTGCATACTTG
CGTCGGTCGGACAAACGAGG

(SEQ ID NO: 1)

The components are arranged in the following order: forward primer (not underlined), secondary read sequence 1 (underlined), mRNA targeting sequence (not underlined), secondary read sequence 2 (underlined), and reverse primer (not underlined). The secondary read sequences are the reverse complement of the corresponding secondary probes. Since only codewords that contained four '1's were used, the primary probes for each mRNA needed to contain 4 different secondary read sequences in this example. However, in order to reduce the overall length of the primary probes, the pool of targeting sequences for each mRNA target was split at random into

two pools. Two secondary read sequences are attached each probe in one of the two pools and the other two secondary read sequences are attached the probes in the other pool. The design criteria for each component are described below.

Primer Design. Specific index primers were generated by a collection of 240,000 published sequences of orthogonal 25-bp long sequences. These sequences were trimmed to 20 bp, selected for a narrow 70 to 80 °C melting temperature, the absence of consecutive repeats of 3 or more base, and the presence of a GC clamp, i.e. one of the two 3' terminal bases must be G or C. To further improve specificity, these sequences were then screened against the human genome using BLAST+ (Camacho et al 2009), and primers with 14 or more contiguous bases of homology were eliminated. In a subsequent screening via BLAST+, primers that shared 11 or more contiguous bases or more than 5 bases at the 3' end of any other primer or the T7 promoter were also removed.

Secondary Probes Design. 30-bp long secondary probe sequences were created by concatenating fragments of the orthogonal primer set described above. These secondaries were then screened for orthogonality with other secondaries (no more than 11 basepairs of homology) and for potential off-target binding sites in the human genome (no more than 14 basepairs of homology). Secondary sequences used in this example are provided in Table 1.

Table 1

<u>Bit</u>	<u>Secondary sequences</u>	<u>Sequence number</u>
B1	CGCAACGCTTGGGACGGTTCCAATCGGATC	SEQ ID NO: 2
B2	CGAATGCTCTGGCCTCGAACGAACGATAGC	SEQ ID NO: 3
B3	ACAAATCCGACCAGATCGGACGATCATGGG	SEQ ID NO: 4
B4	CAAGTATGCAGCGCGATTGACCGTCTCGTT	SEQ ID NO: 5
B5	TGCGTCGTCTGGCTAGCACGGCACGCAAAT	SEQ ID NO: 6
B6	AAGTCGTACGCCGATGCGCAGCAATTCACT	SEQ ID NO: 7
B7	CGAAACATCGGCCACGGTCCCGTTGAACTT	SEQ ID NO: 8
B8	ACGAATCCACCGTCCAGCGCGTCAAACAGA	SEQ ID NO: 9
B9	CGCGAAATCCCCGTAACGAGCGTCCCTTGC	SEQ ID NO: 10

B10	GCATGAGTTGCCTGGCGTTGCGACGACTAA	SEQ ID NO: 11
B11	CCGTCGTCTCCGGTCCACCGTTGCGCTTAC	SEQ ID NO: 12
B12	GGCCAATGGCCCAGGTCCGTCACGCAATTT	SEQ ID NO: 13
B13	TTGATCGAATCGGAGCGTAGCGGAATCTGC	SEQ ID NO: 14
B14	CGCGCGGATCCGCTTGTCGGGAACGGATAC	SEQ ID NO: 15
B15	GCCTCGATTACGACGGATGTAATTCGGCCG	SEQ ID NO: 16
B16	GCCCGTATTCCCGCTTGCGAGTAGGGCAAT	SEQ ID NO: 17

mRNA Targeting Sequence Design. To determine the relative abundance of all the isoforms of all genes expressed in these cell lines, transcriptome profiling data from the ENCODE project for total RNA from A549 and IMR90 cells was processed using the publicly available software *cufflinks*, along with human genome annotations from gencode v18. Gene models corresponding to the most highly expressed isoform were used to build a sequence library in FASTA format recording the dominant isoform of every gene. Genes of interest were selected from this library. These genes were partitioned into 1kb segments, then the software *OligoArray2.1* was used to generate primary probe sequences for the human transcriptome with the following constraints: 30-bp or 40-bp length, depending on the experiment; probe-target melting temperatures greater than 70 °C (variable parameter); no cross hybridization targets with melting temperatures greater than 72 °C (variable parameter); no predicted internal secondary structures with melting temperatures greater than 76 °C (variable parameter); and no single-nucleotide contiguous repeats of 6 or more bases. After OligoArray probe selection, all potential probes that mapped to a different gene were rejected while all potential probes with multiple alignments to the same gene were retained. A BLAST database was assembled from the FASTA library of all expressed genes to screen for probes' uniqueness. For each gene, 14 to 28 targeting sequences produced during the OligoArray processing were selected.

Probe Synthesis – index PCR. The template for specific probe sets were selected from the complex oligopool via limited-cycle PCR. Briefly, 0.5 to 1 ng of the complex oligopool was combined with 0.5 micromolar of each primer. The forward primer matched the priming sequence for the desired subset while the reverse primer was a 5' concatenation of this sequence with a T7 promoter. To avoid the generation of G-quadruplets, which can be difficult to

synthesize, the terminal Gs required in the T7 promoter were generated from Gs located at the 5' of the priming region where appropriate. All primers were synthesized by IDT. A 50 microliter reaction volume was amplified either using the KAPA real-time library amplification kit (KAPA Biosystems; KK2701) or via a homemade qPCR mix which included 0.8X EvaGreen (Biotum; 31000-T) and the hot-start Phusion polymerase (New England Biolabs; M0535S). Amplification was followed in real time using Agilent's MX300P or Biorad's CFX Connect. Individual samples were removed immediately before the plateau in amplification to minimize distortion of template abundance due to over-amplification. Individual templates were purified with columns according to the manufacturer's instructions (Zymo DNA Clean and Concentrator; D4003) and eluted in RNase-free deionized water.

Amplification via *in-vitro* transcription. The template was then amplified via *in vitro* transcription. Briefly, 0.5 to 1 microgram of template DNA was amplified into 100-200 microgram of RNA in a single 20-30 microliter reaction with a high yield RNA polymerase (New England Biolabs; E2040S). Reactions were supplemented with 1X RNase inhibitor (Promega RNasin; N2611). Amplification was typically run for 4 to 16 hours at 37 °C to maximize the yield. The RNA was not purified after the reaction and was either stored at -80 °C or immediately converted into DNA as described below.

Reverse Transcription. 1 to 2 nmol of fluorescently-labeled ssDNA probe was created from the above *in vitro* transcription reactions using the reverse transcriptase Maxima H- (Thermo Scientific; EP0751). This enzyme was used because of its higher processivity and temperature resistance, which allowed the conversion of large quantities of RNA into DNA within small volumes at temperatures that disfavor secondary structure formation. The unpurified RNA created above was supplemented with 1.6 mM of each dNTP, 1-2 nmol of fluorescently labeled forward primer, 300 units of Maxima H-, 60 units of RNasin, and a final 1X concentration of the Maxima RT buffer. The final 75 microliter volume was incubated at 50 °C for 60 minutes.

Strand Selection and Purification. The template RNA in the reaction above was then removed from the DNA via alkaline hydrolysis. 75 microliters of 0.25 M EDTA and 0.5 N NaOH were added to each reverse transcription reaction, and the sample was incubated at 95 °C for 10 minutes. The reaction was immediately neutralized by purifying the ssDNA probe with a modified version of the Zymo Oligo Clean and Concentrator protocol. Specifically, the 5-

microgram capacity column was replaced with a 25-microgram or 100-microgram capacity DNA column as appropriate. The remainder of the protocol was run according to the manufacturer's instructions. Probe was eluted in 100 microliter RNase-free deionized water and evaporated in a vacuum concentrator. The final pellet was resuspended in 10 microliters of RNase-free water and stored at -20 °C. Denaturing poly-acrylimide gel electrophoresis and absorption spectroscopy revealed that this protocol typically produced 90-100% incorporation of the fluorescent primer into full length probe and 75-90% recovery of the total fluorescent probe. Thus, without exceeding a 150-microliter reaction volume, this protocol can be used to create ~2 nmol of fluorescent probe.

Cell culture and fixation. A549 and IMR90 cells (American Type Culture Collection) were cultured with Dulbecco's Modified Eagle Medium and Eagle's Minimum Essential Medium respectively. Cells were incubated at 37 °C with 5% CO₂ for 36-48 hours. Cells were fixed in 3% paraformaldehyde (Electron Microscopy Sciences) in PBS for 15 minutes, washed with PBS, and permeabilized in 70% ethanol overnight at 4 °C.

Fluorescence In Situ Hybridization (FISH) – primary probes. Cells were hydrated in wash buffers (2xSSC, 50% formamide) for 10 minutes, labelled with primary oligos (0.5 nM per sequence) in hybridization buffers (2xSSC, 50% formamide, 1 mg/mL yeast tRNA, and 10% dextran sulfate) overnight at 37 °C, washed with wash buffers at 47 °C for 10 minutes twice, and washed with 2xSSC twice. Fluorescent fiducial beads (Molecular Probes, F-8809) were added at a 1:10,000 dilution in 2xSSC before imaging.

Secondary probes. Secondary probes (10 nM) were hybridized in secondary hybridization buffers (2xSSC, 20% formamide, and 10% dextran sulfate) to their primary targets for 30 minutes at 37 °C. Cells remained on the microscope stage during the hybridization. An objective heater was used to maintain the temperature at 37 °C. Cells were washed with secondary wash buffers (2xSSC, 20% formamide).

Fluidics and STORM Imaging. Multiple rounds of sequential labeling, washing, imaging, and bleaching were performed on an automated platform consisted of a fluidics setup and a STORM (stochastic optical reconstruction microscopy) microscope. The fluidics setup included a flow chamber (Bioptech FCS2), a peristaltic pump (Rainin Dynamax RP-1), and three computer-controlled 8-way valves (Hamilton MVP and Hamilton HVXM 8-5). This system allowed the automated integration of STORM movie collection and secondary hybridization.

The imaging buffer included, 50 mM Tris (pH 8) 10% (w/v) glucose, 1% β ME (2-mercaptoethanol) or 25 mM MEA, with or without 2 mM 1,5-cyclooctadiene, and an oxygen scavenging system (0.5 mg/ml glucose oxidase (Sigma-Aldrich) and 40 microgram/ml catalase (Sigma-Aldrich)). A layer of mineral oil was used to seal the imaging buffer, preventing its acidification over the course of multiple hybridizations.

The STORM setup included an Olympus IX-71 inverted microscope configured for oblique incidence excitation. The samples were continuously illuminated with a 642-nm diode-pumped solid-state laser (VFL-P500-642; MPB communications). A 405-nm solid-state laser (Cube 405-100C; Coherent) was used for activation of dyes. Fluorescence was collected using an Olympus (UPlanSApo 100x, 1.4 NA) objective lens and passed through a custom dichroic, as well as a quad-view beam splitter. All movies were recorded using an EMCCD camera (Andor iAxon 897), imaging at 60Hz. The 512x256 field of view of the camera was split into separate 256x256 pixel movies prior to saving. The left half of this field of view contained the STORM data and the right half contained images of the fluorescent fiducial beads. These latter movies were downsampled to 1 Hz prior to saving. During data acquisition, a home-built focus lock was used to maintain a constant focal plane. STORM movies included 20,000 to 30,000 frames in STORM buffer while the bleach movies included 10,000 frames in wash buffer.

Image Analysis – analysis of single-molecule localizations. Movies of single molecule localizations and fluorescent fiducial beads were processed separately using a previously published single-emitter localization software.

Image Registration. The starting position of the beads from each round of hybridization were used to align movies from each round. The 2D autocorrelation between bead images of consecutive hybridizations followed by nearest-neighbor matching was used to match beads between images. The pair of beads with the most similar displacement vector were used to compute a rigid translation-rotation warp to align the beads. This alignment method is robust to samples in which multiple fiducials are displaced or come detached and reattach during imaging.

Drift Correction. Drift during image acquisition was corrected using the trajectory of the fiducial beads (recorded at 1 Hz). Bead positions were linked in each frame. The trajectory of the two beads that moved in the most correlated fashion was taken as the drift trajectory.

mRNA Cluster Calling. Localizations were first screened to be above a threshold number of photons (generally 2000) and required to be within 32 nm of 5 other localizations (parameters

may be tuned). The remaining molecule localizations were binned in a 2D histogram of 10x10 nm bins (bin size is a variable parameter). All connected bins were taken to be part of a cluster (diagonal contacts are classified as connected). Clusters were required to have more than 80 total localizations across all hybridizations (variable parameter) to be called an mRNA cluster.

5 The weighted centroids of these clusters from the 2D histogram were recorded as the mRNA positions.

A given cluster is recorded to be represented in an individual hybridization round if more than 9 localizations (variable parameter) are found within a 48 nm radius (variable parameter) of the centroid for that mRNA in each hybridization round.

10 Cluster Decoding. For each mRNA cluster, a codeword is readout, including “0”s for all the hybridization rounds in which less than the threshold number of localizations are found near the centroid and “1”s for the rounds where above threshold number of localizations are counted. The SECDED codebook decoded these as either perfect matches to target mRNA codewords, correctable errors which can be unambiguously mapped back to target mRNA, or uncorrectable
15 errors, which differed by two or more letters from the words in the codebook.

Fig. 2A shows a STORM image of a cell. Fig. 2B shows a zoom in of the boxed region in Fig. 2A. Each dot indicates a localization. Localizations from different rounds of imaging are shown differently. Fig. 2C shows a representative cluster of localizations from the boxed region in Fig. 2B. The cluster shows localization signals from 4 different hybridizations. This cluster is
20 a putative mRNA encoded with codeword [0 1 0 1 1 1 0 0]. Fig. 2D shows a reconstructed cell image of 14 genes after decoding and error correction. Each gene is shown differently. Fig. 2E shows measured gene expression for the 14 genes from the cell. Fig. 2F shows a comparison of transcript count with ensemble RNA-sequencing data. Fig. 2G shows correlation of transcript expression level between two cells detected using the described approach.

25

While several embodiments of the present invention have been described and illustrated herein, those of ordinary skill in the art will readily envision a variety of other means and/or structures for performing the functions and/or obtaining the results and/or one or more of the advantages described herein, and each of such variations and/or modifications is deemed to be
30 within the scope of the present invention. More generally, those skilled in the art will readily appreciate that all parameters, dimensions, materials, and configurations described herein are

meant to be exemplary and that the actual parameters, dimensions, materials, and/or configurations will depend upon the specific application or applications for which the teachings of the present invention is/are used. Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific
5 embodiments of the invention described herein. It is, therefore, to be understood that the foregoing embodiments are presented by way of example only and that, within the scope of the appended claims and equivalents thereto, the invention may be practiced otherwise than as specifically described and claimed. The present invention is directed to each individual feature, system, article, material, kit, and/or method described herein. In addition, any combination of
10 two or more such features, systems, articles, materials, kits, and/or methods, if such features, systems, articles, materials, kits, and/or methods are not mutually inconsistent, is included within the scope of the present invention.

All definitions, as defined and used herein, should be understood to control over dictionary definitions, definitions in documents incorporated by reference, and/or ordinary
15 meanings of the defined terms.

The indefinite articles “a” and “an,” as used herein in the specification and in the claims, unless clearly indicated to the contrary, should be understood to mean “at least one.”

The phrase “and/or,” as used herein in the specification and in the claims, should be understood to mean “either or both” of the elements so conjoined, i.e., elements that are
20 conjunctively present in some cases and disjunctively present in other cases. Multiple elements listed with “and/or” should be construed in the same fashion, i.e., “one or more” of the elements so conjoined. Other elements may optionally be present other than the elements specifically identified by the “and/or” clause, whether related or unrelated to those elements specifically identified. Thus, as a non-limiting example, a reference to “A and/or B”, when used in
25 conjunction with open-ended language such as “comprising” can refer, in one embodiment, to A only (optionally including elements other than B); in another embodiment, to B only (optionally including elements other than A); in yet another embodiment, to both A and B (optionally including other elements); etc.

As used herein in the specification and in the claims, “or” should be understood to have
30 the same meaning as “and/or” as defined above. For example, when separating items in a list, “or” or “and/or” shall be interpreted as being inclusive, i.e., the inclusion of at least one, but also

including more than one, of a number or list of elements, and, optionally, additional unlisted items. Only terms clearly indicated to the contrary, such as “only one of” or “exactly one of,” or, when used in the claims, “consisting of,” will refer to the inclusion of exactly one element of a number or list of elements. In general, the term “or” as used herein shall only be interpreted as
5 indicating exclusive alternatives (i.e. “one or the other but not both”) when preceded by terms of exclusivity, such as “either,” “one of,” “only one of,” or “exactly one of.” “Consisting essentially of,” when used in the claims, shall have its ordinary meaning as used in the field of patent law.

As used herein in the specification and in the claims, the phrase “at least one,” in
10 reference to a list of one or more elements, should be understood to mean at least one element selected from any one or more of the elements in the list of elements, but not necessarily including at least one of each and every element specifically listed within the list of elements and not excluding any combinations of elements in the list of elements. This definition also allows that elements may optionally be present other than the elements specifically identified within the
15 list of elements to which the phrase “at least one” refers, whether related or unrelated to those elements specifically identified. Thus, as a non-limiting example, “at least one of A and B” (or, equivalently, “at least one of A or B,” or, equivalently “at least one of A and/or B”) can refer, in one embodiment, to at least one, optionally including more than one, A, with no B present (and optionally including elements other than B); in another embodiment, to at least one, optionally
20 including more than one, B, with no A present (and optionally including elements other than A); in yet another embodiment, to at least one, optionally including more than one, A, and at least one, optionally including more than one, B (and optionally including other elements); etc.

When the word “about” is used herein in reference to a number, it should be understood that still another embodiment of the invention includes that number not modified by the presence
25 of the word “about.”

It should also be understood that, unless clearly indicated to the contrary, in any methods claimed herein that include more than one step or act, the order of the steps or acts of the method is not necessarily limited to the order in which the steps or acts of the method are recited.

In the claims, as well as in the specification above, all transitional phrases such as
30 “comprising,” “including,” “carrying,” “having,” “containing,” “involving,” “holding,” “composed of,” and the like are to be understood to be open-ended, i.e., to mean including but

not limited to. Only the transitional phrases “consisting of” and “consisting essentially of” shall be closed or semi-closed transitional phrases, respectively, as set forth in the United States Patent Office Manual of Patent Examining Procedures, Section 2111.03.

What is claimed is:

CLAIMS

1. A method, comprising:
 - exposing a sample to a plurality of nucleic acid probes;
 - 5 for each of the nucleic acid probes, determining binding of the nucleic acid probes within the sample;
 - creating codewords based on the binding of the nucleic acid probes; and
 - for at least some of the codewords, matching the codeword to a valid codeword wherein, if no match is found, applying error correction to the codeword to form a valid
 - 10 codeword.
2. The method of claim 1, comprising exposing the sample to at least 5 different nucleic acid probes.
- 15 3. The method of any one of claims 1 or 2, comprising exposing the sample to at least 10 different nucleic acid probes.
4. The method of any one of claims 1-3, comprising exposing the sample to at least 100 different nucleic acid probes.
- 20 5. The method of any one of claims 1-4, comprising exposing the sample to a plurality of nucleic acid probes simultaneously.
6. The method of any one of claims 1-5, comprising exposing the sample to a plurality of
- 25 nucleic acid probes sequentially.
7. The method of claims 1-6, wherein the plurality of nucleic acid probes comprises a combinatorial combination of nucleic acid probes with different sequences.
- 30 8. The method of claim 7, wherein the combinatorial combination of nucleic acid probes target a combinatorial combination of RNA species in the sample.

9. The method of claim 7 or 8, wherein the combinatorial combination of nucleic acid probes target a combinatorial combination of DNA sequences in the sample.
- 5 10. The method of any one of claims 1-9, wherein at least some of the plurality of nucleic acid probes comprises a first portion comprising a target sequence and a second portion comprising one or more read sequences.
- 10 11. The method of claim 10, wherein the plurality of nucleic acid probes comprises distinguishable nucleic acid probes formed from combinatorial combination of one or more read sequences taken from the one or more read sequences.
12. The method of any one of claims 10 or 11, wherein the target sequence is substantially complementary to a nucleic acid sequence encoding a protein.
- 15 13. The method of any one of claims 10-12, wherein the target sequence is substantially complementary to messenger RNA (mRNA).
14. The method of any one of claims 10-13, wherein the plurality of nucleic acid probes comprises at least 8 possible read sequences.
- 20 15. The method of any one of claims 10-14, wherein the plurality of nucleic acid probes comprises no more than 32 possible read sequences.
- 25 16. The method of any one of claims 10-15, wherein the plurality of nucleic acid probes comprises no more than 16 possible read sequences.
17. The method of any one of claims 10-16, wherein the plurality of nucleic acid probes comprises no more than 8 possible read sequences.

18. The method of any one of claims 10-17, wherein the plurality of read sequences are distributed on the plurality of nucleic acid probes so as to define an error-correcting code.
19. The method of any one of claims 10-18, wherein the target sequence of the plurality of nucleic acid probes has an average length of between 10 and 200 nucleotides.
20. The method of any one of claims 10-19, wherein the plurality of read sequences have an average length of between 5 nucleotides and 50 nucleotides.
21. The method of any one of claims 10-20, wherein at least some of the plurality of nucleic acid probes comprise no more than 10 read sequences.
22. The method of any one of claims 10-21, wherein at least some of the plurality of nucleic acid probes comprise no more than 6 read sequences.
23. The method of any one of claims 10-22, wherein at least some of the plurality of nucleic acid probes comprise no more than 4 read sequences.
24. The method of any one of claims 10-23, wherein at least some of the plurality of nucleic acid probes comprise no more than 3 read sequences.
25. The method of any one of claims 10-24, wherein at least some of the plurality of nucleic acid probes comprise no more than 2 read sequences.
26. The method of claims 1-25, further comprising exposing the sample to a first secondary probe comprising a first signaling entity, the first secondary probe able to bind to some of the read sequences of the nucleic acid probes, and determining binding of the nucleic acid probes by determining the first signaling entity within the sample.
27. The method of claim 26, wherein the first signaling entity is fluorescent.

28. The method of any one of claims 26 or 27, wherein the first signaling entity is a protein.
29. The method of any one of claims 26-28, wherein the first signaling entity is a dye.
- 5 30. The method of any one of claims 26-29, wherein the first signaling entity is a nanoparticle.
- 10 31. The method of any one of claims 26-30, further comprising exposing the sample to a second secondary probe comprising a second signaling entity, the second secondary probe able to bind to some of read sequences of the nucleic acid probes, and determining binding of the nucleic acid probes by determining the second signaling entity within the sample.
- 15 32. The method of claim 31, wherein the first signaling entity and the second signaling entity are identical.
33. The method of claim 31, wherein the first signaling entity and the second signaling entity are not identical.
- 20 34. The method of any one of claims 31-33, further comprising inactivating the first signaling entity prior to exposing the sample to the second secondary probe.
- 25 35. The method of claim 34, comprising inactivating the first signaling entity by photobleaching at least some of the first signaling entity.
36. The method of any one of claims 34 or 35, comprising inactivating the first signaling entity by chemically bleaching at least some of the first signaling entity.
- 30 37. The method of any one of claims 34-36, comprising inactivating the first signaling entity by exposing the first signaling entity to a reactant able to alter the structure of the signaling entity.

38. The method of any one of claims 34-37, comprising inactivating the first signaling entity by removing at least some of the first signaling entity.
- 5 39. The method of any one of claims 34-38, comprising inactivating the first signaling entity by dissociating the first signaling entity from the first secondary probe.
40. The method of any one of claims 34-39, comprising inactivating the first signaling entity by dissociating the first secondary probe that contains the first signaling entity from the
10 sample.
41. The method of any one of claims 34-40, comprising inactivating the first signaling entity by exposing the signaling entity or the first secondary probe to an enzyme.
- 15 42. The method of claims 1-41, wherein at least some of the plurality of nucleic acid probes comprise DNA.
43. The method of any one of claims 1-42, wherein at least some of the plurality of nucleic acid probes comprise RNA.
20
44. The method of any one of claims 1-43, wherein at least some of the plurality of nucleic acid probes comprise PNA.
45. The method of any one of claims 1-44, wherein at least some of the plurality of nucleic acid probes comprise LNA.
25
46. The method of any one of claims 1-45, wherein the plurality of nucleic acid probes have an average length of between 10 and 300 nucleotides.
- 30 47. The method of any one of claims 1-46, wherein at least some of the plurality of nucleic acids probes are configured to bind to a nucleic acid within the sample.

48. The method of any one of claims 1-47, wherein at least some of the binding of the nucleic acid probes to a target within the sample is specific binding.
- 5 49. The method of any one of claims 1-48, wherein at least some of the binding of the nucleic acid probes to a target within the sample is via Watson-Crick base pairing.
50. The method of any one of claims 1-49, wherein at least some of the plurality of nucleic acids probes are configured to bind to RNA.
- 10 51. The method of any one of claims 1-50, wherein at least some of the plurality of nucleic acid probes are configured to bind to non-coding RNA.
52. The method of any one of claims 1-51, wherein at least some of the plurality of nucleic acid probes are configured to bind to mRNA.
- 15 53. The method of any one of claims 1-52, wherein at least some of the plurality of nucleic acid probes are configured to bind to transfer RNA (tRNA).
- 20 54. The method of any one of claims 1-53, wherein at least some of the plurality of nucleic acid probes are configured to bind to ribosomal RNA (rRNA).
55. The method of any one of claims 1-54, wherein at least some of the plurality of nucleic acid probes are configured to bind to lncRNA.
- 25 56. The method of any one of claims 1-55, wherein at least some of the plurality of nucleic acid probes are configured to bind to snoRNA.
57. The method of any one of claims 1-56, wherein at least some of the plurality of nucleic acid probes are configured to bind to non-coding RNA.
- 30

58. The method of any one of claims 1-57, wherein the plurality of nucleic acids probes is configured to bind to DNA.
59. The method of any one of claims 1-58, wherein at least some of the plurality of nucleic acid probes are configured to bind to genomic DNA.
60. The method of any one of claims 1-59, comprising determining binding of the nucleic acid probes within the sample at a resolution better than 300 nm.
61. The method of any one of claims 1-60, comprising determining binding of the nucleic acid probes within the sample at a resolution better than 100 nm.
62. The method of any one of claims 1-61, comprising determining binding of the nucleic acid probes within the sample at a resolution better than 80 nm.
63. The method of any one of claims 1-62, comprising determining binding of the nucleic acid probes within the sample at a resolution better than 50 nm.
64. The method of any one of claims 1-63, wherein the sample comprises a cell.
65. The method of claim 64, wherein the cell is a human cell.
66. The method of any one of claims 64 or 65, wherein the cell is fixed.
67. The method of any one of claims 1-66, comprising determining binding of the nucleic acid probes by imaging at least a portion of the sample.
68. The method of any one of claims 1-67, comprising determining binding of the nucleic acid probes using an optical imaging technique.

69. The method of any one of claims 1-68, comprising determining binding of the nucleic acid probes using a fluorescence imaging technique.
- 5 70. The method of any one of claims 1-69, comprising determining binding of the nucleic acid probes using a super-resolution fluorescence imaging technique.
71. The method of claim 70, comprising determining binding of the nucleic acid probes using stochastic optical reconstruction microscopy (STORM).
- 10 72. The method of any one of claims 70 or 71, comprising determining binding of the nucleic acid probes using photoactivated localization microscopy (PALM) or fluorescence photoactivation localization microscopy (FPALM).
- 15 73. The method of any one of claims 70-72, comprising determining binding of the nucleic acid probes using stimulated emission depletion microscopy (STED).
74. The method of any one of claims 70-73, comprising determining binding of the nucleic acid probes using structured illumination microscopy (SIM).
- 20 75. The method of any one of claims 70-74, comprising determining binding of the nucleic acid probes using reversible saturated optically linear fluorescence transition (RESOLFT) microscopy.
- 25 76. A method, comprising:
exposing a sample to a plurality of nucleic acid probes, wherein the nucleic acid probes comprise a first portion comprising a target sequence and a second portion comprising one or more read sequences, and wherein at least some of the plurality of nucleic acid probes comprises distinguishable nucleic acid probes formed from combinatorial combination of one or more read sequences taken from a plurality of read
30 sequences; and

for each of the nucleic acid probes, determining binding of the target sequences of the nucleic acid probes within the sample.

77. The method of claim 76, further comprising:

5 creating codewords based on determination of the read sequences within the sample.

78. The method of claim 77, further comprising:

10 for at least some of the codewords, matching the codeword to a valid codeword wherein, if no match is found, applying error correction to the codeword to form a valid codeword.

79. A method, comprising:

15 exposing a sample to a plurality of primary nucleic acid probes;
 exposing the plurality of primary nucleic acid probes to a sequence of secondary nucleic acid probes and determining fluorescence of each of the secondary nucleic acid probes within the sample;

 creating codewords based on fluorescence of the secondary nucleic acid probes;
 and

20 for at least some of the codewords, matching the codeword to a valid codeword wherein, if no match is found, applying error correction to the codeword to form a valid codeword.

80. A method, comprising:

25 exposing a plurality of primary nucleic acid probes to a sample; and
 exposing the plurality of nucleic acid probes to a sequence of secondary nucleic acid probes and determining fluorescence of each of the secondary probes within the sample, wherein at least some of the plurality of secondary nucleic acid probes comprises distinguishable secondary nucleic acid probes formed from combinatorial combination of
30 one or more read sequences taken from a plurality of read sequences.

81. The method of claim 80, further comprising:
creating codewords based on fluorescence of the secondary nucleic acid probes.
82. The method of claim 81, further comprising:
5 for at least some of the codewords, matching the codeword to a valid codeword
wherein, if no match is found, applying error correction to the codeword to form a valid
codeword.
83. A composition, comprising:
10 a plurality of nucleic acid probes, each comprising a first portion comprising a
target sequence and a plurality of read sequences, wherein the plurality of read sequences
are distributed on the plurality of nucleic acid probes so as to define an error-correcting
code.
- 15 84. The composition of claim 83, wherein the plurality of nucleic acid probes defines a code
space with a Hamming distance of at least 2.
85. The composition of claim 84, wherein the plurality of nucleic acid probes defines a code
space with a Hamming distance of at least 3.
20
86. The composition of any one of claims 84 or 85, wherein the code is a Hamming(7, 4)
code, a Hamming(15, 11) code, a Hamming(31, 26) code, a Hamming(63, 57) code, or a
Hamming(127, 120) code.
- 25 87. The composition of any one of claims 83-86, wherein the code is a SECDED code.
88. The composition of claim 87, wherein the code is a SECDED(8,4) code.
89. The composition of claim 87, wherein the code is a SECDED(16,4) code.
30

90. The composition of claim 87, wherein the code is a SECDED(16, 11) code, a SECDED (22, 16) code, a SECDED (39, 32) code, or a SECDED (72, 64) code.
- 5 91. The composition of any one of claims 87-90, wherein only codewords with a constant number of ones are used.
92. The composition of any one of claims 83-91, wherein the plurality of nucleic acid probes comprises no more than 100 possible read sequences.
- 10 93. The composition of any one of claims 83-92, wherein the plurality of nucleic acid probes comprises no more than 64 possible read sequences.
94. The composition of any one of claims 83-93, wherein the plurality of nucleic acid probes comprises no more than 32 possible read sequences.
- 15 95. The composition of any one of claims 83-94, wherein the plurality of nucleic acid probes comprises no more than 16 possible read sequences.
96. The composition of any one of claims 83-95, wherein the plurality of nucleic acid probes comprises no more than 8 possible read sequences.
- 20 97. The composition of any one of claims 83-96, wherein the target sequence of the plurality of nucleic acid probes has an average length of between 10 and 200 nucleotides.
- 25 98. The composition of any one of claims 83-97, wherein the plurality of read sequences have an average length of between 5 nucleotides and 50 nucleotides.
99. The composition of any one of claims 83-98, wherein at least some of the plurality of nucleic acid probes comprise no more than 10 read sequences.

100. The composition of any one of claims 83-99, wherein at least some of the plurality of nucleic acid probes comprise no more than 6 read sequences.
- 5 101. The composition of any one of claims 83-100, wherein at least some of the plurality of nucleic acid probes comprise no more than 4 read sequences.
102. The composition of any one of claims 83-101, wherein at least some of the plurality of nucleic acid probes comprise no more than 2 read sequences.
- 10 103. The composition of any one of claims 83-102, wherein at least some of the plurality of nucleic acid probes comprise DNA.
104. The composition of any one of claims 83-103, wherein at least some of the plurality of nucleic acid probes comprise RNA.
- 15 105. The composition of any one of claims 83-104, wherein at least some of the plurality of nucleic acid probes comprise PNA.
106. The composition of any one of claims 83-105, wherein at least some of the plurality of nucleic acid probes comprise LNA.
- 20 107. The composition of any one of claims 83-106, wherein the plurality of nucleic acid probes have an average length of between 10 and 300 nucleotides.
- 25 108. A method, comprising:
- associating a plurality of targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code space;
- 30 associating a plurality of distinguishable read sequences with the plurality of codewords such that each distinguishable read sequence represents a value of a position

within the codewords; and

forming a plurality of nucleic acid probes, each comprising a target sequence and one or more read sequences.

- 5 109. The method of claim 108, wherein the code space has a Hamming distance of at least 2.
110. The method of any one of claims 108 or 109, wherein the plurality of targets comprise RNA.
- 10 111. A method, comprising:
- exposing nucleic acid probes formed using the method of any one of claims 108-110 to a sample;
- sequentially determining binding of each of a plurality of secondary nucleic probes to the nucleic acid probes within the sample;
- 15 determining sample codewords comprising a pattern of bindings of the plurality of secondary nucleic probes to the nucleic acid probes; and
- matching the sample codewords to the codewords to determining targets within the sample.
- 20 112. The method of claim 111, wherein at least some of the plurality of secondary nucleic probes is able to bind to a read sequence of a nucleic acid probe.
113. The method of claim 112, wherein at least some of the plurality of secondary nucleic probes is able to specifically bind to only one read sequence.
- 25 114. The method of claim 111-113, wherein if no match is found between a codeword and the codeword derived from the sample codeword, applying error correction to the sample codeword to identified a matched codeword.
- 30 115. The method of claim 111-114, wherein at least some of the secondary nucleic acid probes each comprise a signaling entity.

116. The method of claim 115, comprising imaging the signaling entity to determine binding of the plurality of secondary nucleic probes within the sample.
- 5 117. The method of claim 116, comprising imaging the signaling entity using a super-resolution imaging technique.
118. The method of any one of claims 116 or 117, further comprising applying drift correction to the images.
- 10 119. The method of any one of claims 116-118, further comprising applying physical registration to the images.
120. The method of claim 111-119, comprising determining binding of the plurality of secondary nucleic probes within the sample at a resolution better than 100 nm.
- 15 121. A method, comprising:
 associating a plurality of targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code space;
 forming a plurality of nucleic acid probes each comprising a target sequence; and
 forming groups comprising the plurality of nucleic acid probes such that each group of nucleic acid probes corresponds to at least one common value of a position within the codewords.
- 20 122. The method of claim 121, where in the codewords comprise a number of positions that is less than the number of targets.
- 25 123. A method, comprising:
 associating a plurality of targets with a plurality of target sequences and a
- 30

plurality of codewords, wherein the codewords comprise a number of positions that is less than the number of targets, and wherein each codeword is associated with a single target;

5 associating a plurality of distinguishable read sequences with the plurality of codewords such that each distinguishable read sequence represents a value of a position within the codewords; and

forming a plurality of nucleic acid probes, each comprising a target sequence and one or more read sequences.

10 124. The method of claim 123, wherein the codewords form an error-checking space.

125. The method of any one of claims 123 or 124, wherein the codewords form an error-correcting code space.

15 126. A method, comprising:

associating a plurality of targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-correcting code space;

20 forming a plurality of nucleic acid probes each comprising a target sequence; and

forming groups comprising the plurality of nucleic acid probes such that each group of nucleic acid probes correspond to at least one common value of a position within the codewords.

25 127. The method of any one of claims 123-126, wherein the groups are applied to the sample sequentially using fluidic devices.

128. The method of any one of claims 123-127, wherein each group of nucleic acid probes contains a group identification sequence distinguishable from the group identification
30 sequence of the other groups.

129. A method, comprising:

associating a plurality of nucleic acid targets with a plurality of target sequences and a plurality of codewords, wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-

correcting code;

associating unique read sequences with each possible value of each position in the codewords, wherein the read sequences are taken from a set of orthogonal sequences, which have limited homology with one another and with the nucleic acid species in a sample;

forming a plurality of primary nucleic acid probes each comprising a target sequence that uniquely binds to a nucleic acid target and one or more read sequences;

forming a plurality of secondary nucleic acid probes comprising a signaling entity and a sequence that is complementary to one of the read sequences;

exposing a sample to the primary nucleic acid probes such that the nucleic acid probes hybridize to the nucleic acid targets in the sample;

exposing the primary nucleic acid probes in the sample to a secondary nucleic acid probe such that the secondary nucleic acid probe hybridizes to the read sequence on at least some of the primary nucleic acid probes;

imaging the sample; and

repeating the exposing and imaging steps one or more times, using a different secondary nucleic acid probe for at least some of the repetitions.

130. The method of claim 129, wherein the read sequences are taken from a set of orthogonal sequences, which have a homology of less than 15 basepairs with one another and with the nucleic acid species in a sample.

131. The method of any one of claims 129 or 130, comprising combining the images of the sample into a combined image.

132. The method of claim 131, further comprising identifying foreground features corresponding to putative nucleic acid targets in the sample.

133. The method of claim 132, comprising determining the codeword corresponding to each foreground feature by determining binding patterns of the nucleic acid probe pools to the sample.
- 5
134. The method of claim 133, comprising matching the codeword to a valid codeword wherein, if no match is found, applying error correction to the codeword to form a valid codeword that correspond to a nucleic acid target
- 10
135. The method of any one of claims 129-134, comprising imaging the sample at a resolution better than 300 nm.
136. The method of any one of claims 129-135, comprising imaging the sample at a resolution better than 100 nm.
- 15
137. The method of any one of claims 129-136, comprising imaging the sample using a super-resolution fluorescence imaging technique.
138. The method of any one of claims 129-137, comprising imaging the sample using a
- 20
- technique selected from the group consisting of STORM, PALM, FPALM, STED, SIM, RESOLFT, SOFI or SPDM.
139. The method of any one of claims 129-138, where in the codewords comprise a number of positions that is less than the number of targets.
- 25
140. A method, comprising:
- associating a plurality of nucleic acid targets with a plurality of target sequences and a plurality of codewords wherein the codewords comprise a number of positions and values for each position, and the codewords form an error-checking and/or error-
- 30
- correcting code space;
- forming a plurality of nucleic acid probes comprising a signaling entity and a

target sequence that uniquely binds to one of the nucleic acid targets;

grouping the nucleic acid probes into a plurality of probe pools, wherein each of the probe pools corresponds to a specific value of a unique position within the codewords;

- 5 exposing a sample to one of the probe pools;
 imaging the sample; and
 repeating the exposing and imaging steps one or more times, using a different probe pool for at least some of the repetitions.

10 141. The method of claim 140, comprising combining the images of the sample into a combined image.

142. The method of claim 141, further comprising identifying foreground features corresponding to putative nucleic acid targets in the sample.

15 143. The method of claim 142, comprising determining the codeword corresponding to each foreground feature by determining binding patterns of the nucleic acid probe pools to the sample.

20 144. The method of claim 143, comprising matching the determined codeword to a valid codeword wherein, if no match is found, applying error correction to the determined codeword to form a valid codeword that correspond to a nucleic acid target.

25 145. The method of any one of claims 140-144, comprising imaging the sample at a resolution better than 300 nm.

146. The method of any one of claims 140-145, comprising imaging the sample at a resolution better than 100 nm.

30 147. The method of any one of claims 140-146, comprising imaging the sample using a super-resolution fluorescence imaging technique.

148. The method of any one of claims 140-147, comprising imaging the sample using a technique selected from the group consisting of STORM, PALM, FPALM, STED, SIM, RESOLFT, SOFI or SPDM.

5

149. The method of any one of claims 140-148, wherein the codewords comprise a number of positions that is less than the number of targets.

ABSTRACT

The present invention generally relates to systems and methods for imaging or determining nucleic acids, for instance, within cells. In some embodiments, the transcriptome of a cell may be determined. Certain embodiments are directed to determining nucleic acids, such as mRNA, within cells at relatively high resolutions. In some embodiments, a plurality of nucleic acid probes may be applied to a sample, and their binding within the sample determined, e.g., using fluorescence, to determine locations of the nucleic acid probes within the sample. In some embodiments, codewords may be based on the binding of the plurality of nucleic acid probes, and in some cases, the codewords may define an error-correcting code to reduce or prevent misidentification of the nucleic acids. In certain cases, a relatively large number of different targets may be identified using a relatively small number of labels, e.g., by using various combinatorial approaches.

Encoding Scheme

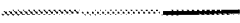
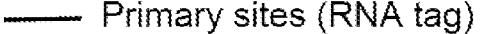
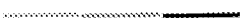
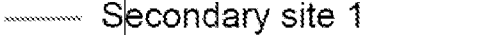
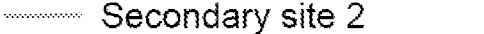
RNA target	codeword	probe sequence	
mRNA I	110		
mRNA II	011		
mRNA III	101		

Fig. 1A

b) Code-Reading Scheme 1: Primaries and Secondaries

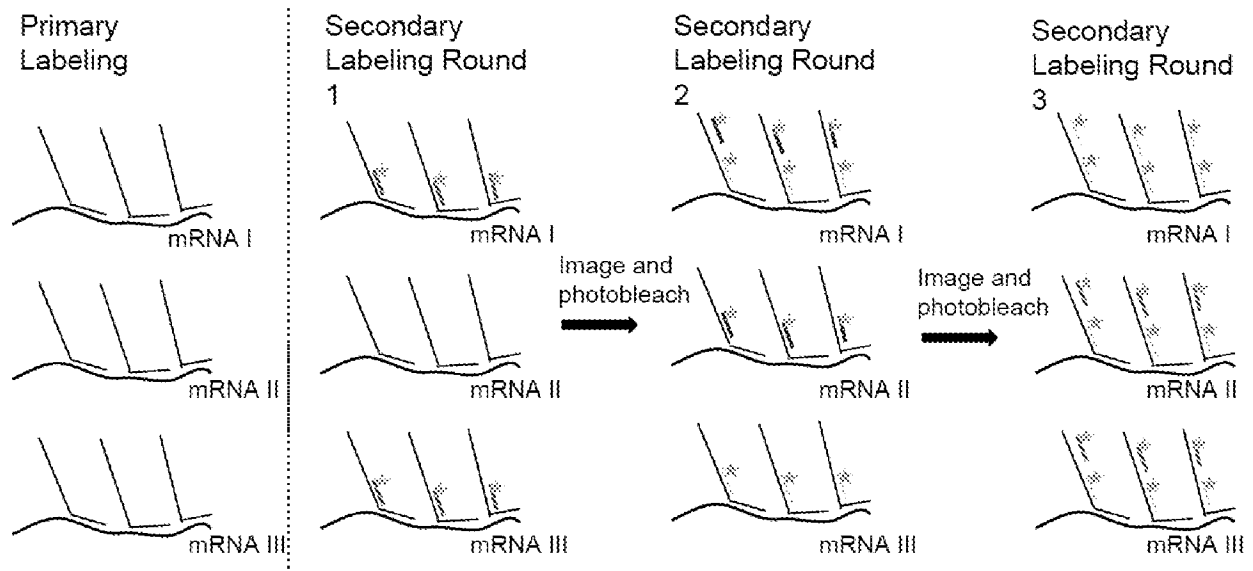


Fig. 1B

Code-Reading Scheme 2: Primaries Only

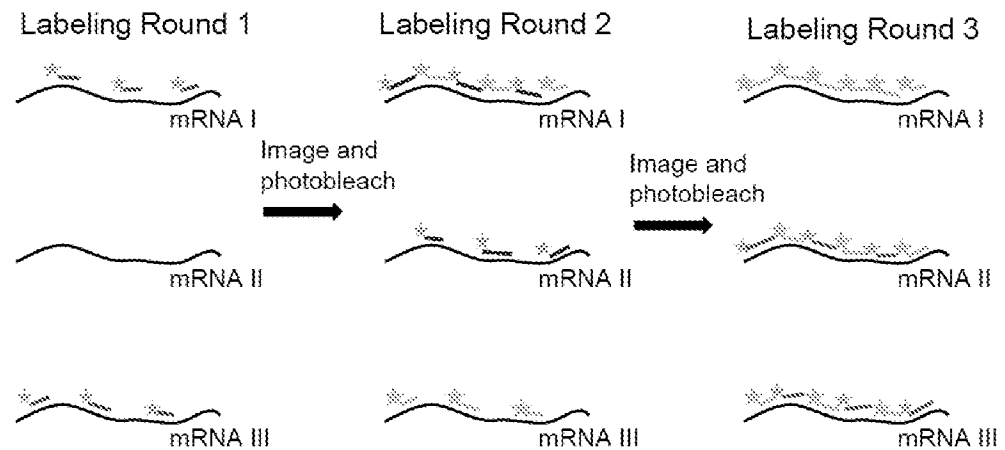


Fig. 1C

Fig. 2A

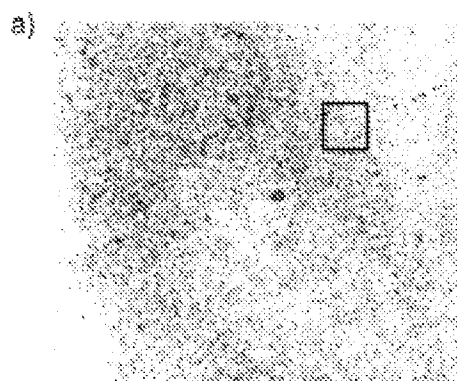


Fig. 2B

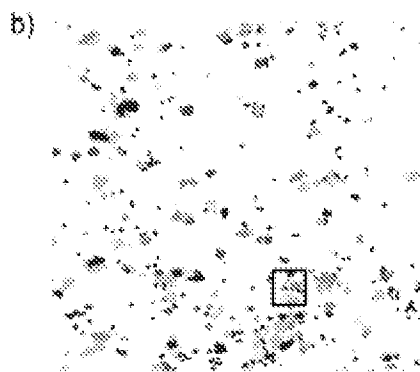


Fig. 2C

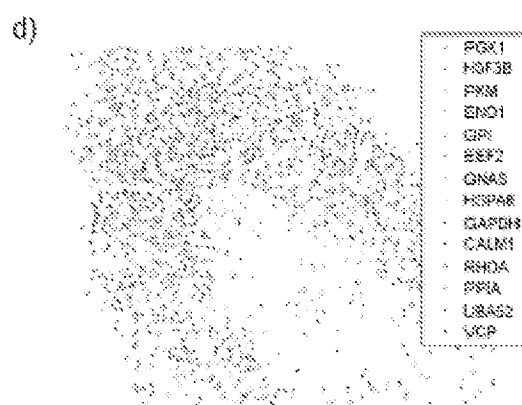
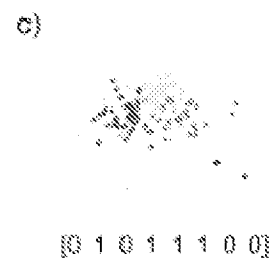


Fig. 2D

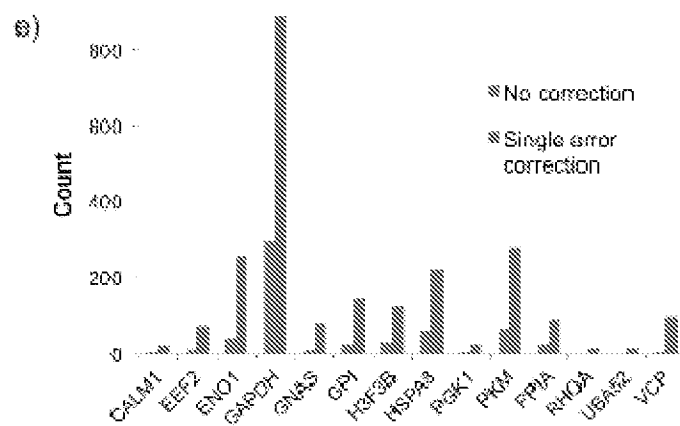


Fig. 2E

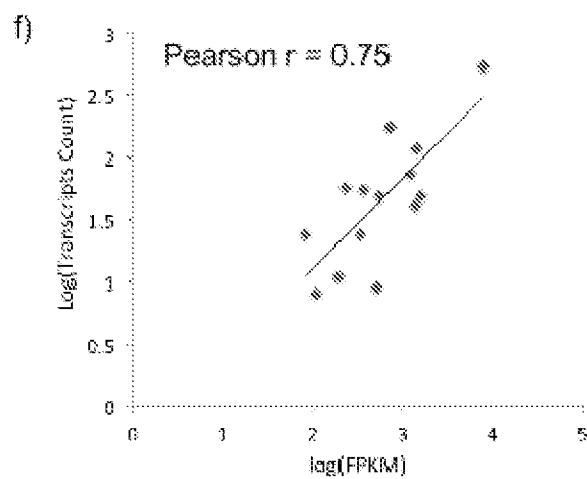


Fig. 2F

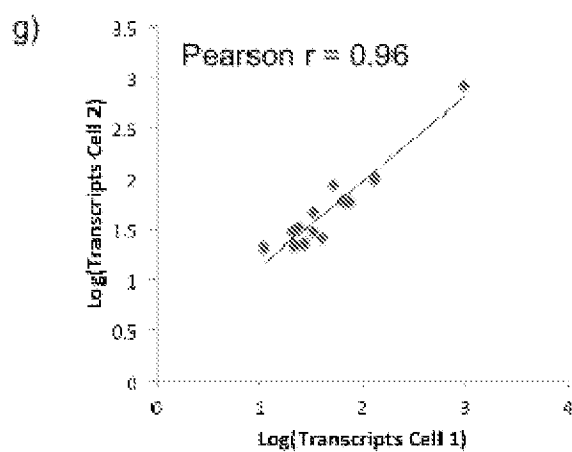


Fig. 2G

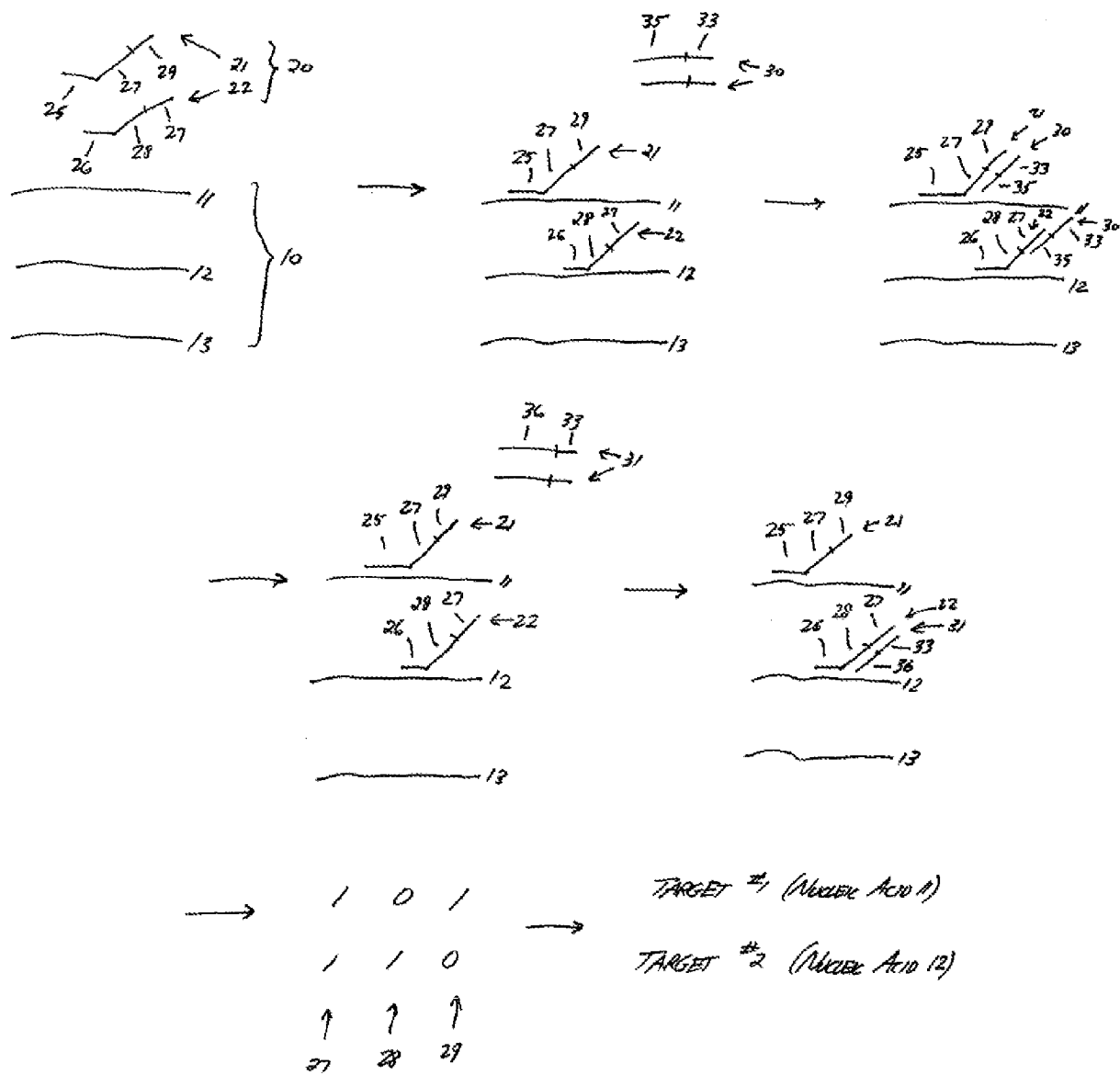


Fig. 3A

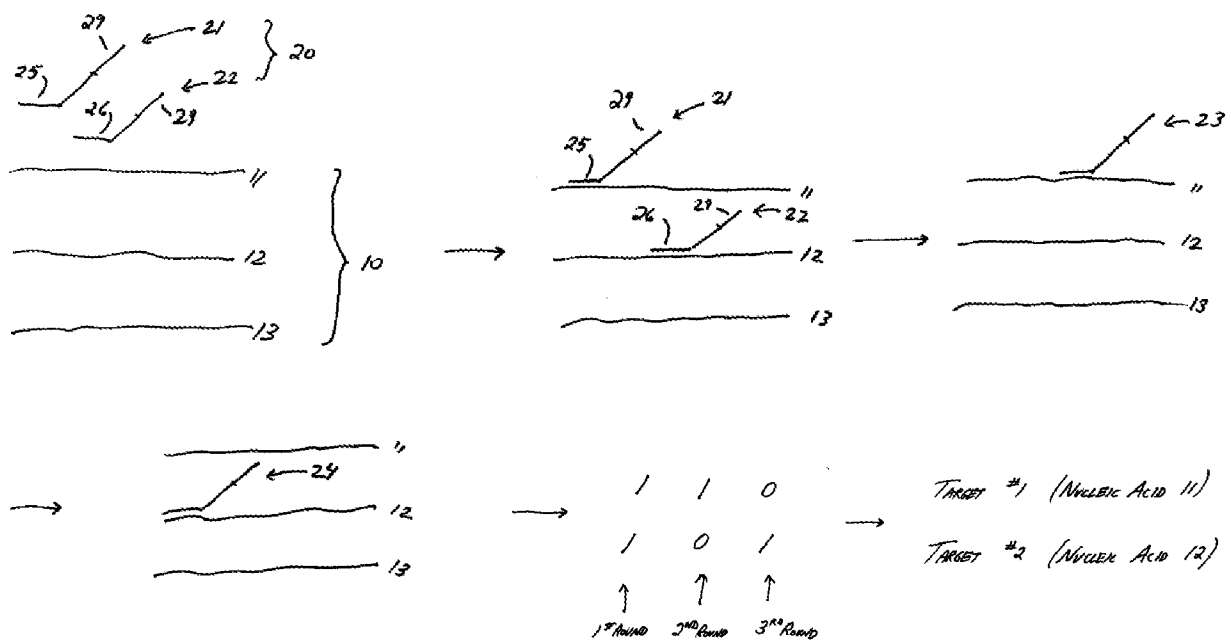


Fig. 3B

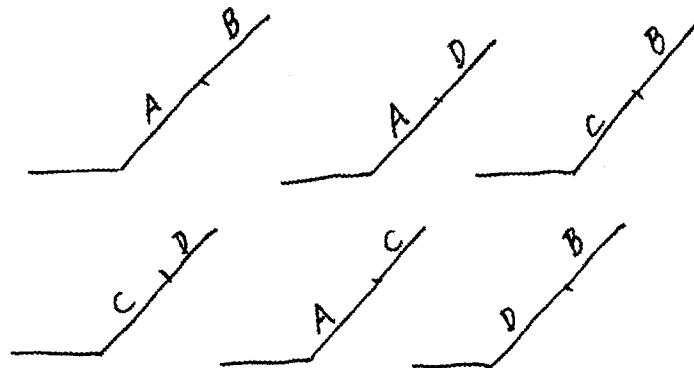


Fig. 4A

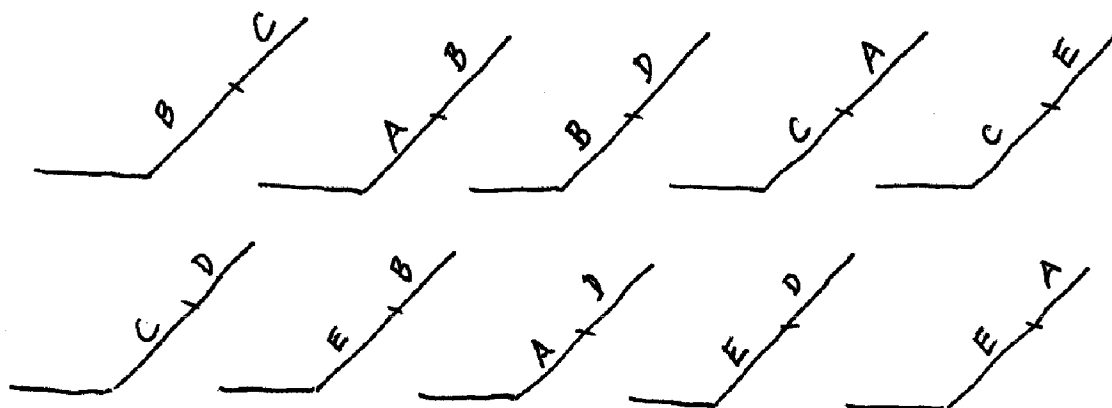


Fig. 4B