

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
12 August 2010 (12.08.2010)(10) International Publication Number
WO 2010/091107 A1(51) International Patent Classification:
H04L 1/00 (2006.01)(21) International Application Number:
PCT/US2010/023083(22) International Filing Date:
3 February 2010 (03.02.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/149,617 3 February 2009 (03.02.2009) US
12/697,995 1 February 2010 (01.02.2010) US(71) Applicant (for all designated States except US): **COMPLETE GENOMICS, INC.** [US/US]; 2071 Stierlin Court, Mountain View, CA 94043 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **KERMANI, Bahram Ghaffarzadeh** [US/US]; 73 Third Street, #12, Los Altos, CA 94022 (US). **SHANNON, Karen, W.** [US/US]; 371 Ocean View Court, Encinitas, CA 92024 (US).(74) Agents: **ALLEN, Kenneth, R.** et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111 (US).

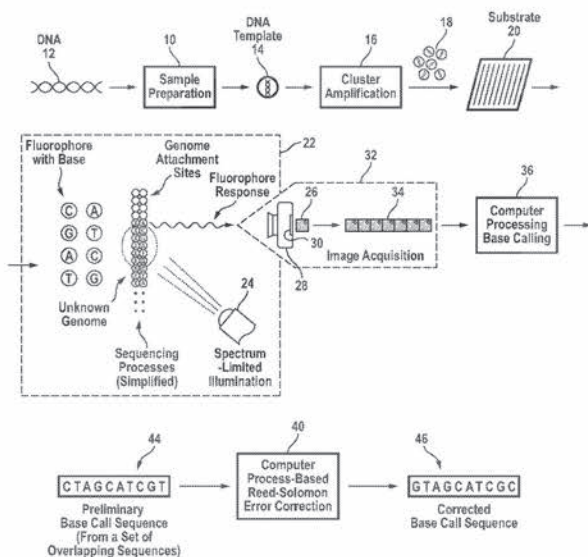
(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG,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

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

Published:

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

(54) Title: METHOD AND APPARATUS FOR QUANTIFICATION OF DNA SEQUENCING QUALITY AND CONSTRUCTION OF A CHARACTERIZABLE MODEL SYSTEM USING REED-SOLOMON CODES



(57) Abstract: Data extracted from fluorophore responses of fluorophore labeled bases in genetic material used in sequencing of unknown fragments from a defined set of for example a model system are converted into a class of block codes that are then employed in a computer-based process to compare and correct preliminary calls of calls of the categorically known genetic material. In a specific embodiment, the Reed-Solomon codes are employed to identify one or more errors as may occur in a finite block of codes. The methodology is also useful to identify elements of a real system containing known elements in the form of a tag. Reed-Solomon sensors may be employed with and in addition to other types of genome sensors.

**METHOD AND APPARATUS FOR QUANTIFICATION OF DNA
SEQUENCING QUALITY AND CONSTRUCTION OF A
CHARACTERIZABLE MODEL SYSTEM USING REED-SOLOMON
CODES**

5

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] The present application claims benefit under 35 USC 119(e) of U.S. provisional Application No. 61/149,617, filed on February 3, 2009, entitled "Method And Apparatus For Correcting DNA Sequencing Errors Using Reed-Solomon Codes."

10

BACKGROUND OF THE INVENTION

[0002] The invention relates to techniques for characterizing the accuracy of genome sequence analysis, and more particularly to the use of advanced mathematical methods for correction of observational errors in sequences having portions of known content.

15 [0003] Large-scale genomic sequence analysis ("sequencing") is a key step toward understanding a wide range of biological phenomena. The need for low-cost, high-throughput sequencing and re-sequencing has led to the development of new approaches to sequencing that employ parallel analysis of multiple nucleic acid targets simultaneously.

[0004] Conventional methods of sequencing are generally restricted to determining a few
20 tens of nucleotides before signals become significantly degraded, thus placing a significant limit on overall sequencing efficiency. Conventional methods of sequencing are also often limited by signal-to-noise ratios that render such methods unsuitable for single-molecule sequencing.

[0005] A challenge of genome sequencing is the accurate recognition, identification,
25 characterization and classification of DNA strands. Efforts have been developed for improving DNA sensing, analysis and measurement throughput by manipulation of DNA including the manipulation of DNA nanoballs ("DNB"). The techniques for sequencing DNB involve the categorization of fluorophore responses of DNB at genome attachment sites on rigid substrates in the presence of interference from adjacent attachment sites. A specific
30 categorization is known as a call, as hereinafter explained. Signal-to-noise ratios in DNA sequencing can be relatively low, which adversely impacts base quality score.

[0006] Improvements in base quality score would allow better characterization of the sequencing system and its failure modes. Improvements would also allow one to quantify improvements in such aspects as the substrate, the biochemistry, the methodology of preparation of samples, the mechanical systems and optical systems, and the mathematical algorithms that analyze and yield the calls.

[0007] Linear block cyclic symbol-based error correction methods relying on error correcting codes have been used to identify and correct bit streams in impaired communication channels, subject to limitations on error rate and run length. Types of codes used in the past for error correction of bit errors in DNA are the Hamming codes. These codes are capable of correcting for one bit error but not one base error. However, Hamming codes are not capable of correcting a large number of errors in a sequence.

[0008] Reed-Solomon error detection/correction is a method based on an error-correcting code that works by oversampling a polynomial constructed from the data. Sampling the polynomial more often than is necessary makes the polynomial over-determined. As long as more than a minimum number of the samples are correct, the original polynomial can be recovered in the presence of a some bad points. The relationship between to good and bad points determines the number of errors that can be corrected.

[0009] Reed-Solomon codes have been explained at length in the mathematics and communication literature. See for example *Error Control Coding: Fundamentals and Applications* by Shu Lin and Daniel Costello; Prentice Hall; and *Error Control Systems for Digital Communication and Storage*, by Stephen B. Wicker; Prentice Hall. It has been shown that if it is guaranteed that there are less than one error in a string of seven values in a sequence having four possible values, then the related mathematics can guarantee that an error in the seven-member long sequence can be captured and corrected.

[0010] It would be advantageous for the field of genome analysis if methods could be designed to characterize and potentially increase the accuracy and call-rate/efficiency of sequencing.

[0011] In the following description, numerous specific details are set forth to provide a more thorough understanding of the present invention. However, it will be apparent to one of skill in the art that the present invention may be practiced without one or more of these specific details. In other instances, well-known features and procedures well known to those skilled in the art have not been described in order to avoid obscuring the invention.

- [0012] The practice of genome analysis may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, and detection of hybridization using a label. The present invention focuses on the detection problem. Conventional techniques and descriptions can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV), *Using Antibodies: A Laboratory Manual*, *Cells: A Laboratory Manual*, *PCR Primer: A Laboratory Manual*, and *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press); Stryer, L. (1995) *Biochemistry* (4th Ed.) Freeman, New York, Gait, "Oligonucleotide Synthesis: A Practical Approach," 1984, IRL Press, London, Nelson and Cox (2000); *Lehninger, Principles of Biochemistry* 3rd Ed., W. H. Freeman Pub., New York, N.Y. and Berg et al. (2002) *Biochemistry*, 5th Ed., W. H. Freeman Pub., New York, N.Y.
- [0013] As used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a polymerase" refers to one agent or mixtures of such agents, and reference to "the method" includes reference to equivalent steps and methods known to those skilled in the art, and so forth.
- [0014] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. All publications mentioned herein are relevant for the purpose of describing and disclosing devices, compositions, formulations and methodologies which are described in the publication and which might be used in connection with the presently described invention.
- [0015] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges is also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either both of those included limits are also included in the invention.

[0016] The following definitions may be helpful in providing background for an understanding of the invention.

[0017] "Adaptor" refers to an engineered construct comprising "adaptor elements" where one or more adaptors may be interspersed within target nucleic acid in a library construct.

5 The adaptor elements or features included in any adaptor vary widely depending on the use of the adaptors, but typically include sites for restriction endonuclease recognition and/or cutting, sites for primer binding (for amplifying the library constructs) or anchor primer binding (for sequencing the target nucleic acids in the library constructs), nickase sites, and the like. In some aspects, adaptors are engineered so as to comprise one or more of the
10 following: 1) a length of about 20 to about 250 nucleotides, or about 40 to about 100 oligonucleotides, or less than about 60 nucleotides, or less than about 50 nucleotides; 2) features so as to be ligated to the target nucleic acid as at least one and typically two "arms"; 3) different and distinct anchor binding sites at the 5' and/or the 3' ends of the adaptor for use in sequencing of adjacent target nucleic acid; and 4) optionally one or more restriction sites

15 [0018] "Amplicon" means the product of a polynucleotide amplification reaction. That is, it is a population of polynucleotides that are replicated from one or more starting sequences. Amplicons may be produced by a variety of amplification reactions, including but not limited to polymerase chain reactions (PCRs), linear polymerase reactions, nucleic acid sequence-based amplification, circle dependant amplification and like reactions (see, e.g., US Pat. Nos.
20 4,683,195; 4,965,188; 4,683,202; 4,800,159; 5,210,015; 6,174,670; 5,399,491; 6,287,824 and 5,854,033; and US Pub. No. 2006/0024711).

[0019] "Circle dependant replication" or "CDR" refers to multiple displacement amplification of a double-stranded circular template using one or more primers annealing to the same strand of the circular template to generate products representing only one strand of
25 the template. In CDR, no additional primer binding sites are generated and the amount of product increases only linearly with time. The primer(s) used may be of a random sequence (e.g., one or more random hexamers) or may have a specific sequence to select for amplification of a desired product. Without further modification of the end product, CDR often results in the creation of a linear construct having multiple copies of a strand of the
30 circular template in tandem, i.e. a linear, concatamer of multiple copies of a strand of the template.

[0020] "Circle dependant amplification" or "CDA" refers to multiple displacement amplification of a double-stranded circular template using primers annealing to both strands

of the circular template to generate products representing both strands of the template, resulting in a cascade of multiple-hybridization, primer-extension and strand-displacement events. This leads to an exponential increase in the number of primer binding sites, with a consequent exponential increase in the amount of product generated over time. The primers
5 used may be of a random sequence (e.g., random hexamers) or may have a specific sequence to select for amplification of a desired product. CDA results in a set of concatameric double-stranded fragments.

[0021] “Complementary” or “substantially complementary” refers to the hybridization or base pairing or the formation of a duplex between nucleotides or nucleic acids, such as, for
10 instance, between the two strands of a double-stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single-stranded nucleic acid. Complementary nucleotides are, generally, A and T (or A and U), or C and G. (Universal bases may be used in some appropriate in some applications.) Two single-stranded RNA or DNA molecules are said to be substantially complementary when the nucleotides of one
15 strand, optimally aligned and compared and with appropriate nucleotide insertions or deletions, pair with at least about 80% of the other strand, usually at least about 90% to about 95%, and even about 98% to about 100%.

[0022] “Duplex” means at least two oligonucleotides or polynucleotides that are fully or partially complementary and which undergo Watson-Crick type base pairing among all or
20 most of their nucleotides so that a stable complex is formed. The terms “annealing” and “hybridization” are used interchangeably to mean formation of a stable duplex. “Perfectly matched” in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double-stranded structure with one another such that every nucleotide in each strand undergoes Watson-Crick base pairing with a nucleotide in the other strand. A
25 “mismatch” in a duplex between two oligonucleotides or polynucleotides means that a pair of nucleotides in the duplex fails to undergo Watson-Crick base pairing.

[0023] “Hybridization” refers to the process in which two single-stranded polynucleotides bind non-covalently to form a stable double-stranded polynucleotide. The resulting (usually) double-stranded polynucleotide is a “hybrid” or “duplex.” “Hybridization conditions” will
30 typically include salt concentrations of less than about 1M, more usually less than about 500 mM and may be less than about 200 mM. A “hybridization buffer” is a buffered salt solution such as 5% SSPE, or other such buffers known in the art. Hybridization temperatures can be as low as 5°C, but are typically greater than 22°C, and more typically greater than about 30°C, and typically in excess of 37°C. Hybridizations are usually performed under stringent

conditions, i.e., conditions under which a probe will hybridize to its target subsequence but will not hybridize to the other, uncomplimentary sequences. Stringent conditions are sequence-dependent and are different in different circumstances. For example, longer fragments may require higher hybridization temperatures for specific hybridization than short fragments. As other factors may affect the stringency of hybridization, including base composition and length of the complementary strands, presence of organic solvents, and the extent of base mismatching, the combination of parameters is more important than the absolute measure of any one parameter alone. Generally stringent conditions are selected to be about 5°C lower than the T_m for the specific sequence at a defined ionic strength and pH. Exemplary stringent conditions include a salt concentration of at least 0.01M to no more than 1M sodium ion concentration (or other salt) at a pH of about 7.0 to about 8.3 and a temperature of at least 25°C. For example, conditions of 5xSSPE (750 mM NaCl, 50 mM sodium phosphate, 5 mM EDTA at pH 7.4) and a temperature of 30°C are suitable for allele-specific probe hybridizations.

[0024] “Isolated” means substantially separated or purified away from contaminants by standard methods. In the case of biological heteropolymers such as polynucleotides (DNA, RNA, etc.) for example, the polynucleotide is substantially separated or purified away from other polynucleotides and other contaminants that are present in the cell of the organism in which the polynucleotide naturally occurs. The term “isolated” also means chemically synthesized or, in the case of a polynucleotide or polypeptide, produced by recombinant expression in a host cell.

[0025] “Ligation” means to form a covalent bond or linkage between the termini of two or more nucleic acids, e.g., oligonucleotides and/or polynucleotides, in a template-driven reaction. The nature of the bond or linkage may vary widely and the ligation may be carried out enzymatically or chemically. As used herein, ligations are usually carried out enzymatically to form a phosphodiester linkage between a 5' carbon terminal nucleotide of one oligonucleotide with a 3' carbon of another nucleotide. Template driven ligation reactions are described in the following references: US Pat. Nos. 4,883,750; 5,476,930; 5,593,826; and 5,871,921.

[0026] “Microarray” or “array” refers to a solid phase support having a surface, preferably but not exclusively a planar or substantially planar surface, which carries an array of sites containing nucleic acids such that each site of the array comprises identical copies of oligonucleotides or polynucleotides and is spatially defined and not substantially overlapping with other member sites of the array; that is, the sites are spatially discrete. The array or

microarray can also comprise a non-planar structure with a surface such as a bead or a well. The oligonucleotides or polynucleotides of the array may be covalently bound to the solid support, or may be non-covalently bound. Conventional microarray technology is reviewed in, e.g., Schena, Ed. (2000), *Microarrays: A Practical Approach* (IRL Press, Oxford). As
5 used herein, “random array” or “random microarray” refers to a microarray where the identity of the oligonucleotides or polynucleotides is not discernable, at least initially, from their location but may be determined by a particular operation on the array, such as by sequencing, hybridizing decoding probes or the like. See, e.g., US Pat. Nos. 6,396,995; 6,544,732; 6,401,267; and 7,070,927; WO publications WO 2006/073504 and 2005/082098; and US Pub
10 Nos. 2007/0207482 and 2007/0087362.

[0027] “Nucleic acid”, “oligonucleotide”, “polynucleotide”, “oligo” or grammatical equivalents used herein refers generally to at least two nucleotides covalently linked together. A nucleic acid generally will contain phosphodiester bonds, although in some cases nucleic acid analogs may be included that have alternative backbones such as phosphoramidite,
15 phosphorodithioate, or methylphosphoroamidite linkages; or peptide nucleic acid backbones and linkages. Other analog nucleic acids include those with bicyclic structures including locked nucleic acids, positive backbones, non-ionic backbones and non-ribose backbones. Modifications of the ribose-phosphate backbone may be done to increase the stability of the molecules; for example, PNA:DNA hybrids can exhibit higher stability in some
20 environments.

[0028] “Preselected,” when used in reference to a block of a monomer subunit sequence that has a coding corresponding to at least one Reed-Solomon code, means that the heteropolymer is designed and/or synthesized to include the block of monomer sequence or that the block of monomer sequence is added to a preexisting heteropolymer sequence. For
25 example, by way of illustration, a polynucleotide can be designed to include a block of five to ten nucleotide bases, such as the seven or ten nucleotide base sequences described herein, or a polynucleotide containing such a five to ten base block can be added to a preexisting polynucleotide by ligation or other standard methods.

[0029] “Primer” means an oligonucleotide, either natural or synthetic, which is capable,
30 upon forming a duplex with a polynucleotide template, of acting as a point of initiation of nucleic acid synthesis and being extended from its 3' end along the template so that an extended duplex is formed. The sequence of nucleotides added during the extension process is determined by the sequence of the template polynucleotide. Primers usually are extended by a DNA polymerase.

[0030] “Probe” means generally an oligonucleotide that is complementary to an oligonucleotide or target nucleic acid under investigation. Probes used in certain aspects of the claimed invention are labeled in a way that permits detection, e.g., with a fluorescent or other optically-discernable tag.

- 5 [0031] “Sequence determination” in reference to a target nucleic acid means determination of information relating to the sequence of nucleotides in the target nucleic acid. Such information may include the identification or determination of partial as well as full sequence information of the target nucleic acid. The sequence information may be determined with varying degrees of statistical reliability or confidence. In one aspect, the term includes the
10 determination of the identity and ordering of a plurality of contiguous nucleotides in a target nucleic acid starting from different nucleotides in the target nucleic acid.

- [0032] “Substrate” refers to a solid phase support having a surface, usually planar or substantially planar, which carries an array of sites for attachment of nucleic acid macromolecules such that each site of the array is spatially defined and not overlapping with
15 other member sites of the array; that is, the sites are spatially discrete and optically resolvable. The nucleic acid macromolecules of the substrates of the invention may be covalently bound to the solid support, or may be non-covalently bound, i.e. through electrostatic forces. Conventional microarray technology is reviewed in, e.g., Schena, Ed. (2000), *Microarrays: A Practical Approach* (IRL Press, Oxford).

- 20 [0033] “Macromolecule” used in relation to a nucleic acid means a nucleic acid having a measurable three dimensional structure, including linear nucleic acid molecules with comprising secondary structures (e.g., amplicons), branched nucleic acid molecules, and multiple separate copies of individual with interacting structural elements, e.g., complementary sequences, palindromes, or other sequence inserts that cause three-
25 dimensional structural elements in the nucleic acid.

[0034] “Target nucleic acid” means a nucleic acid from a gene, a regulatory element, genomic DNA, cDNA, RNAs including mRNAs, rRNAs, siRNAs, miRNAs and the like and fragments thereof. A target nucleic acid may be a nucleic acid from a sample, or a secondary nucleic acid such as a product of an amplification reaction.

- 30 [0035] Although the present invention is described primarily with reference to specific embodiments, it is also envisioned that other embodiments will become apparent to those skilled in the art upon reading the present disclosure, and it is intended that such embodiments be contained within the present inventive methods.

SUMMARY OF THE INVENTION

- [0036] According to the invention, data extracted from labeled bases in genetic material, such as fluorophore responses of fluorophore labeled bases in genetic material, used in sequencing of unknown fragments from a defined set of for example a model system are converted into a class of block codes that are then employed in a computer-based process to compare and correct preliminary calls of the unknown genetic material. In a specific embodiment, the Reed-Solomon codes are employed to identify, detect and preferably correct one or more errors as may occur in a finite block of codes corresponding to a DNA sequence.
- 5 The methodology is not only useful for verification of known sequences of a model system used to characterize a real system, it is useful to identify elements of a real system containing known elements in the form of a tag. Reed-Solomon sensors may be employed with and in addition to other types of genome sensors. Compositions of materials are disclosed that can serve as models and diagnostic tools.
- 10 [0037] In a specific embodiment of the invention, using a seven base model system and more specifically an RS(7,5) code using 256 (7-base) sequences, namely having a block size of seven and information content of five, Reed-Solomon error correction can unambiguously correct one short burst of error, i.e., a single base transition error in a sequence, as well as correct two no-calls. (A no-call is: 1) a base that could not be called; or 2) a base that has a quality score less than a preselected confidence threshold.) As a consequence, processing according to the invention can recover 100% of the valid DNB call sequences that incurred 1 error or less in their 7-base sequence while suppressing many invalid or low quality DNB call sequences.
- 15 [0038] A further specific embodiment and in some applications a preferred embodiment comprises a 10-base model system, namely, an RS(10,8) code using 4096 (10-base) sequences that is capable of greater error correction. It is also to be understood that a 5-base model system may also be usefully employed under selected conditions, namely an RS(5,3) code using 5-base sequences.
- 20 [0039] Sequence listings for two artificial sequences according to the invention are included in the text and in a separate sequence listing. The invention will be better understood by reference to the following detailed description in connection with the accompanying drawings.
- 25 30

BRIEF DESCRIPTION OF THE DRAWINGS

[0040] Figure 1 is a block diagram of an environment that may employ a process according to the invention, which shows preliminary base call sequence = SEQ ID NO:4097 and corrected base call sequence = SEQ ID NO:4098.

- 5 [0041] Figure 2 is a table illustrating experimental results of the use of Reed-Solomon correction on a set of 7-mer DNA sequences.

[0042] Figure 3 is a table summarizing a comparison of representation of a typical set of DNB calls showing improvements according to the invention

- 10 [0043] Figure 4 is a table illustrating experimental results of a Reed-Solomon correction on a set of 10-mer DNA sequences (SEQ ID NOS:4099, 3909, 3762, 3762, 4100, 697, 4101, 1316, 1928, 1928, 1515, 1515, 4102, 2378, 2378, 4016, 4016, 4103, 1708, 4104, 3608, 3970, 3970, 1316, 1316, 4105 and 2346, respectively).

[0044] Figure 5 is a depiction of an apparatus incorporating the invention.

15

DETAILED DESCRIPTION OF THE INVENTION

[0045] A genome is not random. A genome is a form of an oriented linear heteropolymer that contains overlapping structures and duplicates in monomer subunit sequences, all of which are built out of a limited set of base pairs, typically denominated C, G, A and T. Some genomes are circular. Thus one can say that for those, the circle can be opened and an arbitrary point and thus be treated as a linear genome. These characteristics can be exploited in the process of genome sequencing. Other constructs in this category include amino acids, monomer subunits as a sequence of nucleotides and artificial DNA sequences comprising at least one oligonucleotide. All of these constructs may be processed as hereinafter explained.

20

[0046] Referring to Figure 1, the overall method for sequencing target nucleic acids, which has been explained in detail elsewhere, includes sample preparation 10 involving extracting and fragmenting target nucleic acids from a DNA sample 12 to produce fragmented target nucleic acid templates 14 that will generally include one or more adaptors. The target nucleic acid templates 14 are optionally subjected to amplification methods 16 to form nucleic acid nanoballs, herein DNBs 18, which are typically disposed on a surface or substrate 20 for purpose of analysis. The substrate may yield patterned or random arrangements of DNBs. Nucleotide sequencing processes 22 are performed on the nucleic acid nanoballs, typically through sequencing-by-ligation techniques, including combinatorial probe anchor ligation

25

30

(“cPAL”) methods, which are described, for example, in Drmanac et al., “Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanarrays,” *Science* 327:78-81, 2009 (January 1, 2010), as well as in published PCT patent applications WO07/133831, WO06/138257, WO06/138284, WO07/044245, WO08/070352, WO08/058282, WO08/070375; and published U.S. patent applications 2007-0037152 and 2008-0221832. In such methods, known labels, such as specific fragments containing a single molecule of a distinguishable fluorophore, are attached as labels according to well-understood rules to the target nucleic acid templates, then resequence indexed on the same types of DNA strand to provide the basis of overlapping data. The sequencing processes 22 referred to herein are merely representative. In another embodiment, tagging is employed. Other processing techniques known or developed in the art may be employed. Then the collection of DNBs on the substrate is irradiated with radiation 24 to excite the fluorophores sufficient to cause the fluorophores associated with each specific label C, G, A or T to fluoresce at their unique wavelengths, from which a spatial image 26 can be made by a camera 28, on a (standard or time-delay integration TDI) CCD array or a scanner in lieu of a CCD array, or other electronic current/voltage sensing techniques. Other sensing mechanisms, such as impedance change sensors, may also be employed. The irradiation may be spectrum specific to excite only a selected fluorophore at a time, which can then be recorded by the camera, or the input to the camera 30 may be filtered to sense and record only spectrum-specific received fluorescent radiation, or all fluorescent radiation can be sensed and recorded simultaneously on a color LCD array and then later analyzed for spectral content at each interrogation site in which there is a nucleic acid construct. The image acquisition 32 yields a series 34 of images of a plurality of interrogation sites that can be analyzed based on spectrum-specific fluorescence intensity through computer processing of the levels of intensity in a process herein denoted as base calling 36 and explained in greater detail herein below. The cPAL and other sequencing methods can also be used to detect specific sequences, such as including Single Nucleotide Polymorphisms (“SNPs”) in nucleic acid constructs, (which include nucleic acid nanoballs as well as linear and circular nucleic acid templates). The calls, or identification of the sequences of base calls, e.g., base calls 44 (herein shown not as overlapping), may contain errors for reasons evident by the nature of the sequencing procedure. Using the computer process- based Reed-Solomon error correction 40 of the invention, whether in the form of a computer processor perform a Reed-Solomon algorithm or in the form of a comparison mechanism using precomputed expected base call sequences, such as in a look-up table, errors can be identified, “nocall” sequences can be flagged and corrections can be made to yield corrected base call sequences 46. It should be

understood that the magnitude of the sites and structures herein depicted are merely a minute fraction of the magnitude of the sites and structures analyzed on a substrate 20, as they do not easily admit to illustration. For example the substrate may be a photolithographically etched, surface modified (SOM) 25 mm by 75 mm silicon substrate with grid-patterned arrays of about 300-nm spots for DNB binding to increase DNA content per array and improve image information density as compared to random genomic DNA arrays.

[0047] Sequencing probes may be detectably labeled with a wide variety of labels. Although the foregoing and following description is primarily directed to embodiments in which the sequencing probes are labeled with fluorophores, it will be appreciated that similar embodiments utilizing sequencing probes comprising other kinds of labels are encompassed by the present invention. Moreover, the processes according to the invention can be employed with unlabeled structures.

[0048] Multiple cycles of cPAL (whether single, double, triple, etc.) will identify multiple bases in the regions of the target nucleic acid adjacent to the adaptors. (It is possible to employ a single cycle of cPAL to render multiple bases in an alternate design.) In brief, cPAL methods are repeatedly executed for interrogation of multiple bases within a target nucleic acid by cycling anchor probe hybridization and enzymatic ligation reactions with sequencing probe pools designed to detect nucleotides at varying positions removed from the interface between the adaptor and target nucleic acid. In any given cycle, the sequencing probes used are designed such that the identity of one or more of the bases at one or more positions is correlated with the identity of the label attached to that sequencing probe. Once the ligated sequencing probe (and hence the base or bases at the interrogation position or positions are detected, the ligated complex is stripped off of the DNB and a new cycle of adaptor and sequencing probe hybridization and ligation is conducted. By this mechanism, oversampled data are obtainable. Oversampling is done by sequencing more cycles than needed to decode an N-mer.

[0049] Four different fluorophores are typically used to identify a base at an interrogation site within a sequencing probe. Conventionally a single base is queried per hybridization-ligation-detection cycle. However, as will be appreciated, embodiments utilizing 8, 16, 20 and 24 fluorophores or more are also encompassed by the present invention. Increasing the number of fluorophores, or using shades of the same fluorophores, or using a combination of the fluorophores increases the number of bases that can be identified during any one cycle.

[0050] In one exemplary embodiment, a set of 7-mer pools of sequencing probes is employed having the following structures, according to conventional notation:

3'-F1-NNNNNNAp

3'-F2-NNNNNNGp

5 3'-F3-NNNNNNCp

3'-F4-NNNNNNTp

[0051] The "p" represents a phosphate available for ligation and "N" represents degenerate bases. F1-F4 represent four different fluorophores – each fluorophore is thus associated with a particular base, A, G, C or T. This exemplary set of probes would allow detection of the base immediately adjacent to the adaptor upon ligation of the sequencing probe to an anchor probe hybridized to the adaptor. To the extent that the ligase used to ligate the sequencing probe to the anchor probe discriminates for complementarity between the base at the interrogation position of the probe and the base at the detection position of the target nucleic acid, the fluorescent signal that would be detected upon hybridization and ligation of the sequencing probe provides the identity of the base at the detection position of the target nucleic acid. In some embodiments, a set of sequencing probes will comprise three differentially labeled sequencing probes, with a fourth optional sequencing probe left unlabeled.

[0052] After performing a hybridization-ligation-detection cycle, the anchor probe-sequencing probe ligation products are stripped and a new cycle is begun. The accuracy of identification of bases and by extension the number of bases that can be consistently and accurately identified can be increased using the error correction methods described herein.

[0053] Imaging acquisition may be performed using methods known in the art, including the use of commercial imaging packages such as Metamorph (Molecular Devices, Sunnyvale, CA). Data extraction may be performed by a series of binaries written in, *e.g.*, C/C++ and base-calling and read-mapping may be performed by a series of Matlab and Perl scripts.

[0054] In an exemplary embodiment, DNBs disposed on a surface undergo a cycle of cPAL in which the sequencing probes utilized are labeled with four different fluorophores (each corresponding to a particular base at an interrogation position within the probe). In a preferred embodiment, to determine the identity of a base of each DNB disposed on the surface, each field of view ("frame") is imaged with four different wavelengths corresponding to the four fluorescently labeled sequencing probes. All images from each

cycle are saved in digital form in a cycle directory, where the number of images is four times the number of frames (where four fluorophores are used). Cycle image data can then be saved into a directory structure organized for downstream processing.

[0055] In some embodiments, data extraction will rely on two types of image data: bright-field images to demarcate the positions of all DNBs on a surface, and sets of fluorescence images acquired during each sequencing cycle. Data extraction software can be used to identify all objects with the bright-field images and then for each such object, the software can be used to compute an average fluorescence value for each sequencing cycle. For any given cycle, there are four data points corresponding to the four images taken at different wavelengths to query whether that base is an A, G, C or T. These raw data points (also referred to herein as “base calls”) are consolidated, yielding a discontinuous sequencing read for each DNB. This read may contain errors due to the ambiguity of the decision-making process in the base call.

[0056] Therefore, according to the invention, error correction 40 is invoked to yield corrected base calls 46. The error correction processes use a class of linear block cyclic symbol-based error correction methods. The methods are based on the use of run-length limited codes of the type known as the Reed-Solomon codes. Reed-Solomon codes have been explained at length in the mathematics and communication literature. Reference is made to the seminal paper “Polynomial Codes Over Certain Finite Fields,” by I.S. Reed and G. Solomon, *SIAM Journal of Applied Math.*, vol. 8, 1960, pp. 300-304 for the mathematical basis. Further reference is made to textbooks for an explanation of the implementation, such as *Digital Communications: Fundamentals and Applications, Second Edition*, by Bernard Sklar (Prentice-Hall, 2001). A brief tutorial on the use of Reed-Solomon codes based on an explanation by Sklar is provided as follows.

[0057] Reed-Solomon codes are *nonbinary cyclic* codes with symbols made up of m -bit sequences, where m is any positive integer having a value greater than 2. $RS(n, k)$ codes on m -bit symbols exist for all n and k for which

$$0 < k < n < 2^m + 2 \quad (1),$$

where k is the number of data symbols being encoded, and n is the total number of code symbols in the encoded block. For the most conventional Reed-Solomon, or $RS(n, k)$ code,

$$(n, k) = (2^m - 1, 2^m - 1 - 2t) \quad (2)$$

where t is the symbol-error correcting capability of the code, and $n - k = 2t$ is the number of parity symbols. An extended Reed-Solomon code can be made up with $n = 2^m$ or $n = 2^m + 1$, but not any further.

[0058] Reed-Solomon codes achieve the *largest possible* code minimum distance for any linear code with the same encoder input and output block lengths. For nonbinary codes, the distance between two code words is defined as the number of symbols in which the sequences differ. For Reed-Solomon codes, the code minimum distance is given by:

$$d_{\min} = n - k + 1 \quad (3)$$

[0059] The code is capable of correcting any combination of t or fewer errors, where t can be expressed as:

$$t = (d_{\min} - 1)/2 = (n - k)/2 \quad (4)$$

where the center portion of this equation “ x ” means the largest integer not to exceed x . Equation (4) illustrates that for the case of Reed-Solomon codes, correcting t symbol errors requires no more than $2t$ parity symbols. Equation (4) lends itself to the following intuitive reasoning. One may say that the decoder has $n - k$ redundant symbols to “spend,” which is twice the amount of correctable errors. For each error, one redundant symbol is used to locate the error, and another redundant symbol is used to find its correct value. The correction is known as an erasure, which is equivalent to a no-call-correcting capability.

[0060] The erasure or no-call-correcting capability ρ of the code is:

$$\rho = d_{\min} - 1 = n - k \quad (5)$$

[0061] Simultaneous error-correction and erasure-correction capability can be expressed as follows:

$$2\alpha + \gamma < d_{\min} < n - k \quad (6)$$

where α is the number of symbol-error patterns that can be corrected and γ is the number of symbol erasure patterns that can be corrected. An advantage of nonbinary codes such as a Reed-Solomon code can be seen by the following comparison. Consider a binary $(n, k) = (7, 3)$ code. The entire n -tuple space contains $2^n = 2^7 = 128$ n -tuples, of which $2^k = 2^3 = 8$ (or $1/16$ of the n -tuples) are code words. Next, consider a *nonbinary* $(n, k) = (7, 3)$ code where each symbol is composed of $m = 3$ bits. The n -tuple space amounts to $2^{nm} = 2^{21} = 2,097,152$ n -tuples, of which $2^{km} = 2^9 = 512$ (or $1/4096$ of the n -tuples) are code words. When dealing with nonbinary symbols, each made up of m bits, only a small fraction (i.e.,

$2km$ of the large number $2nm$) of possible n -tuples are code words. This fraction decreases with increasing values of m . The important point here is that when a small fraction of the n -tuple space is used for code words, a large $d_{\min}$ can be created.

[0062] Reed-Solomon codes have the remarkable property that they are able to correct *any* set of $n - k$ symbol erasures within the block, erasure being equivalent to a no-call. Reed-Solomon codes can be designed to have any redundancy. However, the complexity of a high-speed implementation increases with redundancy. Thus, the more attractive Reed-Solomon codes have high code rates (low redundancy).

[0063] Referring now a specific implementation of the invention, Reed-Solomon error detection and Reed-Solomon error correction is based on selection and use of a specific error-correcting code that in general works by oversampling a polynomial constructed from the data. In the present context, the coefficients of the polynomial are the base call designations from fluorophore-label identification in the base calling analysis coded as integers, e.g., 1,2,3,4, which has been made redundant in nature by the overlapping probes, where the redundancy comes from the length of the probes. As noted, sampling the polynomial more often than is necessary makes the polynomial over-determined. So long as more than a minimum number of the samples is correct, the original polynomial can be recovered in the presence of some bad points. As noted, the relationship between good and bad points determines the number of errors that can be corrected, which for k bases, with $n - k$ redundant bases, have the following properties: where $2t = n - k$ and the location of the error is not known, t errors can be identified and corrected. Where location of the errors is known, then $2t$ errors can be corrected. In other words, Reed-Solomon methodology can correct half as many errors as there are redundant symbols in a block when location of errors is not known and if location is known, it can correct as many errors as there are redundant symbols. The method according to the invention relies on generating a correction code set wherein each random pair from the code set is guaranteed to have a minimum ratio R of distance D to length L of greater than 20 percent from every other member of the set. The distance D greater than or equal to 3 for length L equal to 10, namely a 10base code. Similarly the distance D greater than or equal to 3 for length L equal to 7, namely a 7base code.

[0064] Location of errors can be identified by base reference to the base calling quality score. the base calling quality score can be obtained directly from commercially available base calling algorithms. A paper describing one commercially available technique was newly published in the aforementioned Drmanac et al., "Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanarrays," *Science*, Vol. 327, pp. 78-

81, January 1, 2010 (also found online at www.scienceexpress.org, dated 5 November 2009, Page 1 (10.1126/science.1181498.)

[0065] Significantly, since Reed-Solomon block codes can correct block errors, it can therefore be used to correct for errors in base identification, unlike Hamming block codes, which can correct only for bit errors.

[0066] For DNA reading applications, it is typically thought desirable that the value of “ n ” be kept small, since the value of “ n ” ultimately defines the number of bases in the final code, and the value needs to be small to minimize the cost of the assay. Further, the probability of a base read cannot be so high that the sequencing technology cannot be enabled. In the most efficient case, the value of t , the number of errors that is to be corrected in any one iteration, is set to “1”. In a further case, 10-base codes may be provided that are able to correct 2 errors. Thus, the artificial 10-mer set that can correct for two errors is as follows:

CAGCAAGCAG (SEQ ID NO:1551)

GGAGAAAACG (SEQ ID NO:9)

15 CGGGAACAAA (SEQ ID NO:257)

CACCCACAGC (SEQ ID NO:266)

AACAGACGAC (SEQ ID NO:2342)

TTTAACATGC (SEQ ID NO:3128)

TCTCACCAGT (SEQ ID NO:319)

20 CCAACCCAGA (SEQ ID NO:325)

AGCCCCAGGG (SEQ ID NO:2131)

GCCGCCCCGAC (SEQ ID NO:2396)

CGCAGCAAAG (SEQ ID NO:94)

ACACGCCGAA (SEQ ID NO:2408)

25 TGGGTCTTAC (SEQ ID NO:3955)

CAAAAGCAAG (SEQ ID NO:377)

GAGAAGAACA (SEQ ID NO:127)

TGTAAGAGCT (SEQ ID NO:2182)

ACCACGCACG (SEQ ID NO:393)

CGACCGAGCA (SEQ ID NO:2197)

GGGCCGCGAG (SEQ ID NO:2459)

ACGGCGAGGC (SEQ ID NO:2204)

5 TGTGCGGCAT (SEQ ID NO:1698)

CTGGGGCTAT (SEQ ID NO:3494)

TCCCTGT TAG (SEQ ID NO:4023)

TTCTATCAGC (SEQ ID NO:456)

TTGTCTGCAG (SEQ ID NO:1761)

10 CGTTGTCTAG (SEQ ID NO:3563).

[0067] The hardware basis for implementing Reed-Solomon codes has been detailed elsewhere for communication applications. Reference is made to "Reed-Solomon Error Correction," by C.K.P. Clarke, *BBC Research & Development White Paper WHP 031*, July 2002, (British Broadcasting Corporation), attached hereto as Appendix A. This white paper describes the principles of a general purpose hardware implementation of an error correcting device using Reed-Solomon coded data as input. In the present context, once a polynomial has been constructed from the DNA base call data, it is processed by circuitry such as that disclosed by Clarke to recover the DNA tags.

[0068] It should be understood that the implementation of this invention is only practical in the context of use of a dedicated or a general purpose computer processor. In addition to a hardware implementation of the Reed-Solomon decoder as described in Appendix A, a Reed-Solomon-code-based decoder may also be implemented through a computer program of a general-purpose digital computer. Representative source code that is in MATLAB source code form (MATLAB 7.9.0/529 (R2009b) for representative Reed-Solomon code designs is set forth below. This includes a 10 10base code design with one error correction capability and a 7base R-S decoder:

10 10base code design

[0069] v.m = 4; % Number of bits per symbol

```

v.n = 2^v.m-1; % Codeword length

v.k = 13;

v.n0 = 10;

v.s = v.n - v.n0;

5   v.k0 = v.k - v.s;

v.t = (v.n-v.k)/2; % Error-correction capability of the code

% Coder/Decoder Construction

coder = fec.rsenc(v.n, v.k);

10  coder.ShortenedLength = v.s;

decoder = fec.rsdec(coder);

%% Construct full data set

v.data0 = fullfact(4*ones(1,v.k0));

15  v.nwords0 = size(v.data0,1); % number of words to process

%% Construct message

v.send0 = encode(coder, v.data0)'; % Encode the data.

%% Select the valid codes

20  v.valid = all(ismember(v.send0,[1 2 3 4]),2);

v.data = v.data0(v.valid,:);

v.nwords = size(v.data,1);

v.send = v.send0(v.valid,:);

v.seq = int2nt(v.send);

25

```

Reed-Solomon Decoder (7base)

```

[0070] function [decoded_integer, decoded0_integer, status, recovery] =
        reed_solomon_decoder_7base(sequence_corrupted_integer, erasures)

5         v.m = 3;    % Number of bits per symbol

        v.n = 2^v.m-1; % Codeword length

        v.k = 5;    % message length

        v.t = (v.n-v.k)/2; % Error-correction capability of the code

10        % Coder/Decoder Construction

        coder = fec.rsenc(v.n, v.k);

        decoder = fec.rsdec(coder);

        u.seq_corrupted_integer = double(sequence_corrupted_integer);

15        u.erasures = erasures;

        [dummy, errors, seq_corrected_integer] = decode(decoder,
        u.seq_corrupted_integer',          u.erasures'); % Decode the noisy code.

20        u.seq_corrected_integer = seq_corrected_integer';

        u.errors = errors';

        recovery = (u.errors==1);

        u.status = all(ismember(u.seq_corrected_integer,[1:4]),2) &
                    ismember(u.errors,[0:v.t]);

25

        decoded0_integer = u.seq_corrected_integer;

```

```

decoded_integer = decoded0_integer;

status = u.status;

decoded_integer(status==0,:) = 0;

```

- 5 [0071] The foregoing source code is applicable to each dataset below, including the following dataset for Reed-Solomon 6 + 4 base, 26 sequence, 2 base error correction, 4 nocall correction (SEQ ID NOS:1551, 9, 257, 266, 2342, 3128, 319, 325, 2131, 2396, 94, 2408, 3955, 377, 127, 2182, 393, 2197, 2459, 2204, 1698, 3494, 4023, 456, 1761, and 3563, respectively):

```

10 CAGCAAGCAG GGAGAAAACG CGGGAACAAA CACCCACAGC AACAGACGAC
    TTTAACATGC TCTCACCAGT CCAACCCAGA AGCCCCAGGG GCCGCCCGAC
    CGCAGCAAAG ACACGCCGAA TGGGTCTTAC CAAAAGCAAG GAGAAGAACA
    TGTAAGAGCT ACCACGCACG CGACCGAGCA GGGCCGCGAG ACGGCGAGGC
    TGTGCGGCAT CTGGGGCTAT TCCCTGTTAG TTCTATCAGC TTGTCTGCAG
15 CGTTGTCTAG

```

- [0072] The foregoing code set is also applicable to the following dataset for Reed-Solomon 8+2 base, 4096 sequence, 1 base error, 2 nocall correction (SEQ ID NOS:1-4096):

```

    TAAAAAAAGG ATGAAAAACC CCACAAAAGC AGACAAAATT CTACAAAATA
    ACCCAAAAGA ATCCAAAATC TGTCAAAATA GGAGAAAACG AGGGAAAACA
20 TGGGAAAAG CTGGAAAAC TGTGAAAAT ACTTAAACC ACAACAAAGC
    CGAACAAATT ATAACAAATA CCCACAAAGA CTCACAAATC GTGACAAATG
    TTTACAAATT GTCCCAAACG CTGCCAAACC GGTCCAAAC CCAGCAAAGT
    AGAGCAAATC CGCGCAAATA ATCGCAAATT TAATCAAAT GAGTCAAAT
    TCGTCAAATT AATTCAAAA TATTCAAACG GCTTCAAATG AGAAGAAACG
25 TGAAGAAAAA GGGAGAAACA AGTAGAAAAT CTTAGAAAAA GCACGAAAGT
    GGCCGAAATA TTAGGAAAAC TGC GGAAAAT CTCGGAAACG TAGGGAAAGA
    GTGGGAAACC CGTGGAAAAC CAATGAAATA AACTGAAATC CCGTGAAAAA
    CCCATAAACT GATATAAATC TCTATAAACC CAACTAAAAT CCGCTAAATT
    AGGCTAAAGC CATCTAAACG CCAGTAAACA ACCGTAAACC AAATTAAAGG
30 GTATTAAAC GGCTTAAAT GAGTTAAAGA TTGTAAACC AAAACAAGC

```

GTAAACAAAG CACAACAAGA CGGAACAAAT ATGAACAAAA TTGAACAACG
 GTTAACAAC TCCACAATG CAAGACAAGT AGGGACAAAC TCATACAAAT
 TAGTACAATT GCGTACAAC TATTACAATG ACTTACAAAA TCTTACAACG
 TCAACCAAGG CAACCCAAGC AACCCCAAGA AGGCCCAAAT CTGCCCAAAA
 5 GGTCCCAACA TGAGCCAATG CTTGCCAATG GTATCCAAGT GGCTCCAAGC
 CGGTCCAAGG AATTCCAACC TGAAGCAACC CGCAGCAAAG TTCAGCAACT
 GGGAGCAAAC CTTAGCAACC TTGCGCAATT GTTCGCAATG GAAGGCAAGC
 ATAGGCAAAG TTAGGCAACA GTGGGCAAAA CGTGGCAACA ATTGGCAACT
 TACTGCAATG GCCTGCAACG CCGTGCAACC ACAATCAAAT AAGATCAATT
 10 ACTATCAACG TCTATCAAAA AACCTCAACT TGGCTCAAGG GGTCTCAAGT
 CCAGTCAAAC GACGTCAATG ACCGTCAAAA TCCGTCAACG CAGGTCAATC
 TAATTCAAGC GTATTCAACA ATGTTCAACG TTGTTCAAAA AAAAAGAACG
 TAAAAGAAAA GAGAAGAACA TCGAAGAATA TTGAAGAAGC AATAAGAAAT
 GACCAGAATA ACCCAGAAAG TCCCAGAACA TACGAGAAAT TGGGAGAAGA
 15 CATGAGAAAC CGATAGAATA ATATAGAATT CCCTAGAAGT AGCTAGAATC
 TCTTAGAAGC TTTTAGAATA GAAACGAATC TCAACGAACC CCCACGAAAG
 GCGACGAAAC CAACCGAACG TACCCGAAAC CATCCGAAAT CCTGCGAACA
 GGCTCGAACG CGGTGCAACC AATTCGAAGG GTTTCGAAAC TGAAGGAAGG
 CTTAGGAAGG AAACGGAATT ACGCGGAAAT TATCGGAATA GCTCGGAACA
 20 GAAGGGAACG TCAGGGAATG AAGGGGAACA TAGGGGAAAG GATGGGAAAT
 GCCTGGAAGC GTCTGGAATA CCGTGGAAGG TTAATGAATT TGCATGAATC
 ACTATGAAGC CGTATGAATT ATTATGAATA GGAAGTAAAC GTCCTGAAAT
 TGGCTGAACC TCCGTGAAGC TTCGTGAATA CCTGTGAAGT AGTGTGAATC
 AAATTGAAAA TAATTGAACG GCATTGAATG CACTTGAAAC GAGTTGAAAG
 25 ACGTTGAATA ATGTTGAAGC TATTTGAAAT GCTAATAAGA GTTAATAATC
 AGACATAAAA TGACATAACG CTACATAAAT CGCCATAAAC GGGCATAAAG
 TGTCATAAAT CTTACATAACG GGTGATAATA GAATATAAAC ATATATAAGG
 CCCTATAATG TAGTATAACC GCGTATAATC GTGTATAAGA CGAACTAAAA
 ATAACTAAAT AGCACTAAAC ATTACTAACG TTTACTAAAA CTAGCTAAAC
 30 CGCGCTAAAT ATCGCTAAAA TTCGCTAACG CAGGCTAAGA TGTGCTAAAC
 TAATCTAATA GCATCTAACA ACGTCTAACG TCGTCTAAAA AATTCTAATT
 TGAAGTAATT TCCAGTAAGA TTCAGTAATC AGTAGTAATA CTTAGTAATT
 AAACGTAAGG GTACGTAAAC GGCCGTAAAT GAGCGTAAGA TTGCGTAACC
 TCAGGTAAGT TGCGGTAATA ACTGGTAAGA ATTGGTAATC CAATGTAAAT
 35 CCGTGTAATT AGGTGTAAGC CATTGTAACG TTAATTAAGG AAGATTAACC

CGTATTAAGG CAACTTAATA AACCTTAATC CCGCTTAAAA CAGGTTAACT
CCTGTTAATG GCATTTAAGT GGCTTTAATA CAAAAACAGA AACAAACAGC
GTCAAACAAG CTGAAACAAC GGTAACACC TTACAACATG CGTCAACATG
CACGAACAGT CGGGAACAAA ATGGAACAAT TATGAACAGC GTTGAACACA
5 TCCTAACAAT CCTTAACAAC TCCACACAGG AGTACACATG AAACCACAGA
CACCCACAGC GAGCCACAGG ATGCCACAAC TATCCACAGT TGCGCACATG
ACTGCACAGG GGATCACAGC GTCTCACAGT CATTCACACA CGAAGACAAG
TTAAGACACT TGCAGACACC GTGAGACAAT CGTAGACACT ATTAGACACA
TTTAGACAAG TGGCGACATC GACGGACAGC ATCGGACAAG TTCGGACACA
10 CAGGGACAGG AGTGGACACC TAATGACATG GCATGACACG ACGTGACACA
TCGTGACAAG GCTTGACAAT ACCATACAAT AACTTACACT AATCTACAAG
TATCTACACA GCTCTACATA GTTCTACAGC GAAGTACATG ACAGTACAAA
TCAGTACACG CCCGTACAAC AAGGTACATA GCGGTACAAG TCTGTACAAT
TACTTACAGC GTCTTACACA CATTTACAGT TACAACCAGG AGGAACCACT
15 CTGAACCACA GGTAACCAAA ACACACCAGA ATACACCATC CCCCACCAGC
AGCCACCATT CTCCACCATA GCGCACCAGG TCTCACCAGT GGCGACCACG
CGGGACCACC AATGACCAGG GTTGACCAAC CTTACCACA CCAACCCAGA
CTAACCCATC ACCACCCAGC CGCACCCATT ATCACCCATA TGTACCCATC
GTACCCACG CGGCCCCACT ATGCCCCACA TTGCCCCAAG GTTCCCCAAT
20 CGAGCCCATA ATAGCCCATT CCCGCCAGT AGCGCCCATC TCTGCCCAGC
TTTGCCCATA TACTCCCAAT TGGTCCCAGA CATTCCCAAC AGCAGCCACG
TGCAGCCAAA TAGAGCCAGT GATAGCCAGG ATTAGCCAAC GGACGCCATA
GCCC GCCAGT AGGCGCCATG TGAGGCCAAT CTAGGCCACG TTCGGCCAAC
GGGGGCCACT AGTGGCCAAA TGTGGCCACG CTTGGCCAAT AAATGCCATC
25 CACTGCCATA ACGTGCCAAC CCAATCCACT CCTATCCAAG CACCTCCAAT
CGGCTCCAGA ATGCTCCAGT TATCTCCAAC ACAGTCCACC CCCGTCCACA
GGATTCCAAT AACTTCCAGG GTCTTCCAAC TGGTTCCACT CTGTTCCAAG
GGTTTCCACG CAAAAGCAAG TACAAGCACC GCCAAGCATC GTCAAGCAGA
CCGAAGCATG CATAAGCACT TCACAGCAAC CCCCAGCACG TAGCAGCATC
30 GCGCAGCACC GTAGAGCAGT GGCAGCAGC CGGGAGCAGG AATGAGCACC
TGATAGCATG CTTTAGCATG ACCACGCACG TCCACGCAAA AAACCGCAAG
TAACCGCACA GCACCGCATA GTACCGCAGC GGCCCGCAGT GAGCCGCAAA
ACGCCGCATG AATCCGCACT TCAGCGCAAT TAGGCGCATT GCGGCGCACT
GATGCGCATG ACTGCGCAAA TCTGCGCACG CAATCGCAGT AGGTCGCAAC
35 CGAAGGCAGA ATAAGGCAGT CCCAGGCATT AGCAGGCAGC GATAGGCACC

TCTAGGCATC TTTAGGCAGA CCGCGGCACT CATCGGCATG CCAGGGCATA
CTAGGGCAGC ACCGGGCATC CGCGGGCAGT ATCGGGCAGA CAGGGGCAAA
TGTGGGCAGC TCGTGGCAGA TTGTGGCATC AGAATGCATG GGGATGCATA
TAACTGCAGT GGCCTGCACA AATCTGCAGA ACAGTGCAGG CTCGTGCATG
5 GCGGTGCAGA GTGGTGCATC CAATTGCACA AACTTGCACC GGTTCGCAGC
CCGAATCAGT AGGAATCATC CGACATCACA ATACATCACT AGCCATCACC
GATCATCAGC ATTCATCAAG TTTCATCACA GTAGATCATG ACGGATCAGC
CGGGATCATT ATGGATCATA TGATATCAGT GACTATCACA TCCTATCATA
TTCTATCAGC AGTTATCAGA CTTTATCAGT AGAACTCACA TGAATCAAG
10 CTAATCACT CGCACTCACC GGGACTCAG TGTACTCACT CTTACTCAAG
GGCCCTCATG ACGCCTCAGT CGGCCTCATC ATAGCTCACC AGCGCTCACT
CTCGCTCACA TAGGCTCAGG GATGCTCAGT CAATCTCATG GCCTCTCAAC
CCGTCTCAAG ATAAGTCATG CCCAGTCAGG GCGAGTCAGC GTGAGTCATA
TACCGTCAGC GTCCGTCACA CATCGTCAGT CGCGGTCATG GGGGGTCATC
15 AAATGTCACT AATTGTCAAG TATTGTCACA GCTTGTCTATA GTTGTCTCAGC
CCAATTCATC AGAATTCAGT CTAATTCAGA GACATTCAAG ACCATTCTATA
ATCATTCTCAGC CAGATTCAAC TGTATTCTAGA TAACTTCATG GCACTTCACG
ACGCTTCACA TCGCTTCAAG GCTCTTCAAT ACAGTTCATT CGAGTTCAGC
AGCGTTCAGA CTCGTTCTAGT AAGGTTCAAT GATGTTCTACA TCTGTTCTATA
20 TTTGTTCTCAGC TGGTTTCTATC GAAAAAGACA TCAAAAGATA TTAAGAGAGC
TGCAAAGAGT AAGAAAGACG TAGAAAGAAA ACTAAAGATT CGTAAAGAGC
AACCAAGATG GCCCAAGAAA TGAGAAGAGA GACGAAGACT TCCGAAGATT
CCTGAAGATC AGTGAAGAGT CTTGAAGAGA CGGTAAGATA ATGTAAGATT
GCAACAGAAC CACACAGATG GAGACAGATC TCGACAGACC GACCCAGACC
25 TCCCCAGATC TTCCCAGAGA CAGCCAGACG CCTCCAGATT AGTCCAGAGC
CGATCAGACC AGCTCAGACA TGCTCAGAAG CTCTCAGACT GATTCAGAGA
TTTTCAGACC AACAGAGACT TGGAGAGAGG GGTAGAGAGT ACACGAGAAT
AAGCGAGATT ACTCGAGACG TCTCGAGAAA AAAGGAGACA TAAGGAGAAG
CACGGAGACC GAGGGAGACG TCGGGAGATG TATGGAGACT GCTGGAGATT
30 CCATGAGAGG ATCTGAGATG TTGATAGATT GTTATAGATG TGACTAGACC
CGCCTAGAAG TTCCTAGACT GGGCTAGAAC CTTCTAGACC GAATTAGAAG
ACATTAGATA ATATTAGAGC CCCTTAGATC AGCTTAGAGT CTCTTAGAGA
AAGTTAGAAA TAGTTAGACG GCGTTAGATG GATTTAGACT TCTTTAGATT
GAAAACGAAC ATAAACGAGG CCAACGATG TAGAACGACC GCGAACGATC
35 GTGAACGAGA TACCACGATC GCCCACGACC CCGCACGACG CATCACGATT

CGCGACGAGG GGGGACGAGC GCTTACGAGA GTTTACGATC TAAACCGATA
GCAACCGACA ACGACCGACG TCGACCGAAA AATACCGATT CTACCCGAGG
GACCCCGAAA ACCCCCGATG GGGCCCGAGT TGTCCCGAGG TACGCCGATT
GCCGCCGACT CATGCCGATC CGATCCGAAA ATATCCGAAT AGCTCCGAAC
5 ATTTCCGACG TTTTCCGAAA CAAAGCGAAT CCGAGCGATT AGGAGCGAGC
CATAGCGACG CCCCGCGACT GATCGCGATC TCTCGCGACC AAAGGCGAAC
GTAGGCGAGG CACGGCGAAA ACGGGCGATC CGGGGCGAGT ATGGGCGAGA
TGATGCGATT TCCTGCGAGA TTCTGCGATC AGTTGCGATA CTTTGCGATT
GCAATCGAGT GGCATCGATA AGACTCGACG TGA CTGAAA GGGCTCGACA
10 AGTCTCGAAT CTTCTCGAAA GGAGTCGATT GCCGTCGAGA GTCGTCGATC
CTGGTCGATG TTATTGAGG AAGTTCGACC CGTTTCGAGG ATAAAGGACC
AGCAAGGACT CTCAAGGACA TAGAAGGAGG GATAAGGAGT GCCCAGGAGG
CCGCAGGAGC AGGCAGGATT CTGCAGGATA AGAGAGGACA TGAGAGGAAG
CTAGAGGACT CGCGAGGACC GGGGAGGACG TGTGAGGACT CTTGAGGAAG
15 GTAACGGATG ACGACGGAGC CGGACGGATT ATGACGGATA CTACCGGACC
CGCCCGGACT ATCCCGGACA TTCCCGGAAG TGTCCGGACC CCGGCGGAGT
AGGGCGGATC GAATCGGACT TCATCGGATT TGCTCGGAGA TAGTCGGAAT
GATTCGGAAG ACTTCGGATA ATTTGCGAGC GGAAGGGACA TACAGGGAGT
AGGAGGGACG TGGAGGGAAA CATAGGGAGC AGCCGGGATG GCGCGGGAGT
20 TCTCGGGAGG TAAGGGGAGA GTAGGGGACC GGCGGGGACT TTGGGGGAAC
AATGGGGAGT CCATGGGAAA ACCTGGGAAC CAGTGGGATA GCTATGGAAC
CCACTGGATT AGACTGGAGC CGCCTGGAGA ATCCTGGAGT CAGCTGGAAT
CCGGTGGACA GAATTGGAGA TTATTGGACC TGCTTGGACT CTCTTGGAAG
AAGTTGGAGG GTGTTGGAAC CGTTTGGACC TCAAATGAAT TAGAATGATT
25 GCGAATGACT GATAATGATG ACTAATGAAA TCTAATGACG GGACATGAGA
TACCATGACT GCCCATGATT AGGCATGAGG CATCATGACC ACCGATGACG
TCCGATGAAA AAATATGAGC GTATATGAAG CACTATGAGA CGGTATGAAT
ATGTATGAAA TTGTATGACG GTTTATGACT GTA ACTGAGT GGC ACTGAGC
CGGACTGAGG AATACTGACC CCTCCTGAAA CAAGCTGAAG TACGCTGACC
30 GCCGCTGATC GTCGCTGAGA CCGGCTGATG CATGCTGACT TCATCTGAGG
TACAGTGATG GCCAGTGACG CCGAGTGACC GAACGTGAAG ACACGTGATA
ATACGTGAGC CCCCGTGATC AGCCGTGAGT CTCCGTGAGA AAGCGTGAAA
TAGCGTGACG GCGCGTGATG GATCGTGACT TCTCGTGATT ACGGGTGACT
AATGGTGATG GCTGGTGAAA TGATGTGACC CGCTGTGAAG TTCTGTGACT
35 GGGTGTGAAC CTTTGTGACC TAAATTGAGC GTAATTGACA ATGATTGACG

TTGATTGAAA CCACTTGAGG ATCCTTGATG GGAGTTGACC GTCGTTGACT
TGGGTTGAAC CATGTTGAGA ACATTTGAAT AAGTTTGATT ACTTTTGACG
TCTTTTGAAA GCAAAATAGA GTAAAATATC GGCAAATATT ACGAAATAGG
TGACAATAAT CTACAATACG TTCCAATAAC GGGCAATACT AGTCAATAAA
5 TGTCAATACG CTTCAATAAT GGAGAATATA GCCGAATAGT AGGGAATATG
GATTAATAAC ATTTAATAGG ATAACATACG TTAACATAAA TAGACATAGC
GTGACATACA CGTACATAAA ATTACATAAT GCCCCATAGC GTCCCATATA
CCGCCATAGG TGAGCATAAC CGCGCATACG TTCGCATAAT GGGGCATACC
CTTGCATAAC AAATCATATT ACGTCATAAT TATTCATATA GCTTCATACA
10 AGAAGATATA CTAAGATATT ACCAGATAGT CGCAGATATC GGGAGATATG
TGTAGATATT GGCCGATACG CGGCGATACC AATCGATAGG GTTCGATAAC
ACAGGATAGA ATAGGATATC CCCGGATAGC AGCGGATATT CTCGGATATA
GCGGGATAGG TCTGGATAGT CAATGATACG TACTGATAAC CATTGATAAT
CGAATATAGG GTGATATAGT TTTATATAGG CATCTATATA CCAGTATATG
15 GACGTATAAC ATCGTATAGG CAGGTATAAG TCGTTATAGG GCTTTATAGT
TCGAACTAGC TTGAACTATA GACCACTAGC ATCCACTAAG TTCCACTACA
CAGCACTAGG AGTCACTACC TGGGACTATC ACATACTATT CGATACTAGC
AGCTACTAGA CTCTACTAGT AAGTACTAAT GATTACTACA TCTTACTATA
TTTTACTAGC GAAACCTAGA TTAACCTACC TGCACCTACT CTCACCTAAG
20 AAGACCTAGG GTGACCTAAC CGTACCTACC TGGCCCTATT GGTCCCTATG
AGAGCCTAAG TGAGCCTACA GACGCCTAGT GGGGCCTAAA AGTGCCTACT
CTTGCCTACA GCTTCCTAAC CCTAGCTAGG CACCGCTAGT CGGCGCTAAA
ATGCGCTAAT TATCGCTAGC GTTCGCTACA TTAGGCTATG CGTGGCTATG
GGATGCTAGT AACTGCTAAG TACTGCTACA GCCTGCTATA GTCTGCTAGC
25 CTGTGCTAGG GAAATCTACT TCAATCTATT TGCATCTAGA TAGATCTAAT
GATATCTAAG ACTATCTATA ATTATCTAGC GCCCTCTAAT TGAGTCTAGT
GACGTCTACA TCCGTCTATA TTCGTCTAGC AGTGTCTAGA CTTGTCTAGT
GTATTCTATG ACGTTCTAGC CGGTTCTATT ATGTTCTATA AAAAAGTATA
GCAAAGTAAG CACAAGTATC GAGAAGTATG ACGAAGTAAA TCGAAGTACG
30 TATAAGTATT GCTAAGTACT GACCAGTACG TCCCAGTATG CAGCAGTACC
AGTCAGTAGG AACGAGTATT CGATAGTACG TTATAGTAAT TGCTAGTAAC
GTGTAGTACT CGTTAGTAAT ATTTAGTAAA TTTTAGTACG TTAACGTAGG
AAGACGTACC CGTACGTAGG CAACCGTATA AACCCGTATC CCGCCGTAAA
CAGGCGTACT CCTGCGTATG GCATCGTAGT GGCTCGTATA GACAGGTATT
35 TCCAGGTACT CCTAGGTACC TAACGGTAAT GAGCGGTACT TCGCGGTATT

AATCGGTAAA TATCGGTACG GCTCGGTATG GAAGGGTATA ACAGGGTAAG
TCAGGGTACA AAGGGGTATG GCGGGGTAAA ACTGGGTACT GTCTGGTACG
CTGTGGTACC GGTGTTAAC CGAATGTAAA ATAATGTAAT AGCATGTAAAC
ATTATGTACG TTTATGTAAA CTAGTGTAAAC CGCGTGTAAT ATCGTGTA
5 TTCGTGTACG CAGGTGTAGA TGTGTGTAAAC TAATTGTATA GCATTGTACA
ACGTTGTACG TCGTTGTAAA AATTTGTATT GGCAATTAAA GAGAATTAGT
TATAATTAGG TGACATTATA TCCCATTAGT CCTCATTAGC AGTCATTATT
CTTCATTATA GGAGATTAAT AACGATTAGG GTCGATTAAC TGGGATTACT
CTGGATTAAG GGTGATTACG ACATATTACC CCCTATTACA TTAACCTATT
10 TGCACCTATC ACTACTTAGC CGTACTTATT ATTACTTATA GGACCTTAAC
GTCCCTTAAT TGGCCTTACC TCCGCTTAGC TTCGCTTATA CCTGCTTAGT
AGTGCTTATC AAATCTTAAA TAATCTTACG GCATCTTATG CACTCTTAAC
GAGTCTTAAG ACGTCTTATA ATGTCTTAGC TATTCTTAAT AGAAGTTAAT
CTAAGTTAAA GACAGTTAGG ATCAGTTAAC CAGAGTTAGC AGTAGTTACG
15 TGTAGTTAAA TCGCGTTAGG GCTCGTTAGT CGAGGTAAAC AGCGGTTAAA
TGCGGTTACG CTCGGTTAAT AAGGGTTAGT TTTGGTTAAC CATTGTTATA
GAAATTTATC TCAATTTACC CCCATTTAAG GCGATTTAAC CAACTTTACG
TACCTTTAAC CATCTTTAAT CCTGTTTACA GGCTTTTACG CGGTTTTACC
AATTTTTAGG GTTTTTTAAAC TTA AAAACTG CGTAAAACTG CAACAAACGA
20 AACCAAACGC GTCCAAACAG CTGCAAACAC GGTCAAACCC CCTGAAACGG
TAATAAACCT GCATAAACTT GGCTAAACGA GAGTAAACAT AATTAAACCA
TATTAAACAG AAAACAACGA CACACAACGC GAGACAACGG ATGACAACAC
TATACAACGT TCCCAACCG AGTCCAACCT GGAGCAACAG AACGCAACGT
AGGGCAACAA TGGGCAACCG CTGGCAACAT GGTGCAACCT ACTTCAACAC
25 TGGAGAACTC CGACGAACAG TTACGAACCT TGCCGAACCC GTGCGAACAT
CGTCGAACCT ATTCGAACCA TTTCGAACAG TCGGGAACGC TTGGGAACTA
CCATGAACTC AGATGAACGT CTATGAACGA GACTGAACAG ACCTGAACTA
ATCTGAACGC CAGTGAACAC TGTTGAACGA AAAATAACCT AATATAACAG
TATATAACCA GCTATAACTA GTTATAACGC ACCCTAACAT CAAGTAACCC
30 AACGTAACCA TACGTAACAG GGTGTAACGA ATATTAACCT CCCTTAACGG
GCGTTAACGC GTGTTAACCT ACAAACACGA ATAAACACTC CCCAACACGC
AGCAACACTT CTCAACACTA GCGAACACGG TCTAACACGT TACCACACGG
AGGCACACCT CTGCACACCA GGTCACACAA AGAGACACTA CTAGACACTT
ACCGACACGT CGCGACACTC GGGGACACTG TGTGACACTT TTGTACACGT
35 AATTACACAC GTTTACACGG GTAACCACCG CGGACCACCT ATGACCACCA

TTGACCACAG GTTACCACAT CCACCCACGA CTACCCACTC ACCCCCACGC
CGCCCCACTT ATCCCCACTA TGTCCTACTC AGGGCCACCC CATGCCACGG
GAATCCACTT TCATCCACCT GCGTCCACAT ACTTCCACCA TCTTCCACAG
GGAAGCACTA GCCAGCACGT AGGAGCACTG AGCCGCACCG TGCCGCACAA
5 TAGCGCACGT GATCGCACGG ATTCGCACAC GCAGGCACGA GTAGGCACTC
GGCGGCACTT ACGGGCACGG TTCTGCACGG CAGTGCACCA CACATCACAT
CGGATCACGA ATGATCACGT TATATCACAC CCACTCACCT CCTCTCACAG
CAAGTCACAA AACGTACACAC GTCGTACGG CCGGTCACTA CTGGTACGC
TCATTACGA TTATTACTC TGCTTCACTT ACTTTCACGT CGTTTCACTC
10 TCAAAGACAC CCAAGACCG TAGAAGACTC GCGAAGACCC CAACAGACAG
TACCAGACCC GCCCAGACTC GTCCAGACGA CCGCAGACTG CATCAGACCT
CCTGAGACAA AAATAGACGT GGCTAGACAG CGGTAGACAC TATTAGACGA
GTTTAGACCC AAAACGACAG TAAACGACCA GCAACGACTA GTAACGACGC
GGCACGACGT GAGACGACAA ACGACGACTG AATACGACCT ACCCCGACCG
15 TCCCCGACAA GGAGCGACGA TACGCGACCT GCCGCGACTT AGGGCGACGG
CATGCGACCC ATTTGACTG CCGAGGACCT CATAGGACTG CGACGGACGA
ATACGGACGT CCCC GGACTT AGCCGGACGC GATCGGACCC TCTCGGACTC
TTTCGGACGA ACGGGGACCC TGATGGACCT CTATGGACAG GACTGGACGA
TTCTGGACCC GGGTGGACAT AGTTGGACCA TGTTGGACAG CTTTGGACCT
20 TAAATGACGT GGCATGACCA AATATGACGA AGACTGACTG GGGCTGACTA
GGAGTGACCT TACGTGACGA GTCGTGACCC TGGGTGACAT CTGGTGACCG
GGTGTGACAG ACATTGACAC CCCTTGACAA AAGTTGACTC CGAAATACCA
ATAAATACCT AGCAATACCC GATAATACGC ATTAATACAG TTTAATACCA
CCGCATACGT AGGCATACTC CTAGATACCC CGCGATACCT ATCGATACCA
25 TTCGATACAG TGTGATACCC AAATATACTG GCATATACAA GAGTATACTA
ACGTATACAG TCGTATACCA GGCATACTG ACGACTACGT CGGACTACTC
AGACCTACCA TGACCTACAG CTACCTACCT CGCCCTACCC GGGCCTACCG
TGTCCTACCT CTTCTACAG GCCGCTACGG CCGGCTACGC AGGGCTACTT
CTGGCTACTA GAATCTACCC TCATCTACTC TTATCTACGA TAGTCTACAC
30 CGTTCTACGA ATTTCTACGT TACAGTACGC GTCAGTACCA CATAGTACGT
ATACGTACTG CCCC GTACGG GCGCGTACGC GTGCGTACTA GTAGGTACCT
GGCGGTACCC CGGGGTACCG TTGGGTACAT AATGGTACGC GTTGGTACAG
ACCTGTACAT TAAATTACTG GCAATTACCG ACGATTACCA TCGATTACAG
GCTATTACAT CCACTTACTC AGACTTACGT CTACTTACGA GACCTTACAG
35 ACCCTTACTA ATCCTTACGC CAGCTTACAC TGTCTTACGA CGATTTACAG

TTATTTACCT TGCTTTACCC GTGTTTACAT CGTTTTACCT ATTTTACCA
 TTTTTTACAG CCAAACCGC AGAAAACCTT CTAAAACCTA ACCAAACCGA
 ATCAAACCTC TGTAACCTA TAACAACCGG ATGCAACCCC ACAGAACCGT
 CGAGAACCTC AGCGAACCTA CTCGAACCTT TCTGAACCGA TTTGAACCTC
 5 TGGTAACCGC CATTAACCAA GTCACACCCG CTGACACCCC GGTACACCAC
 ACACCACCGC CGACCACCTT ATACCACCTA CCCCCACCGA CTCCCACCTC
 GTGCCACCTG TTTCCACCTT CGGGCACCCA ATGGCACCCCT GTTGCACCAA
 GACTCACCTT TCCTCACCTT CCTTCACCCC GCAAGACCGT GGCAGACCTA
 AGACGACCCG TGACGACCAA GGGCGACCCA AGTCGACCAT CTTCGACCAA
 10 GGAGGACCTT GCCGGACCGA GTCGGACCTC CTGGGACCTG TTATGACCGG
 AAGTGACCCC CGTTGACCGG CAAATACCAT CCGATACCTT AGGATACCGC
 CATATACCCG CCCCTACCTT GATCTACCTC TCTCTACCCC AAAGTACCAC
 GTAGTACCGG CACGTACCAA ACGGTACCTC CGGGTACCGT ATGGTACCGA
 TGATTACCTT TCCTTACCGA TTCTTACCTC AGTTTACCTA CTTTACCTT
 15 TTCAACCCTG AAACACCCGC GTACACCCAG CACCACCCGA CGGCACCCAT
 ATGCACCCAA TTGCACCCCG GTTCACCCCT ATTGACCCTG GGATACCCGA
 TACTACCCCT GCCTACCTT AGGTACCCGG CATTACCCCC CAAACCCCGC
 AACACCCCGA AGGACCCCAT CTGACCCCAA GTACCCCCA TCACCCCGG
 AAAGCCCCGT GGCGCCCCAG CGGGCCCCAC TATGCCCCGA GTTGCCCCC
 20 CCTTCCCCAA TTGAGCCCTT GTTAGCCCTG TGACGCCCCC CGCCGCCAG
 TTCCGCCCTT GGGCGCCAC CTTCGCCCC GAATGCCAG ACATGCCCTA
 ATATGCCCGC CCCTGCCCTC AGCTGCCCGT CTCTGCCCGA AAGTGCCCAA
 TAGTGCCCGG GCGTGCCCTG GATTGCCCTT TCTTGCCCTT AACATCCCTT
 TGGATCCCGG GGTATCCCGT AACTCCCAT AAGCTCCCTT ACTCTCCCCG
 25 TCTCTCCCAA AAAGTCCCA TAAGTCCCAG CACGTCCCCC GAGGTCCCCG
 TCGGTCCCTG TATGTCCCCT GCTGTCCCTT CCATTCCCGG ATCTTCCCTG
 GACAAGCCTA ACCAAGCCAG TCCAAGCCCA AAACAGCCCG TAACAGCCAA
 GAGCAGCCCA TCGCAGCCTA TTGCAGCCGC AATCAGCCAT GAAGAGCCTT
 TCAGAGCCCT GCGGAGCCAT ACTGAGCCCA TCTGAGCCAG AGGTAGCCCC
 30 CATTAGCCGG CAAACGCCCG TACACGCCAC CATA CGCCAT GAACCGCCTC
 TCACCGCCCC CCCCCGCCAG GCGCCGCCAC TTGGCGCCGT AATGCGCCAC
 GTTGCGCCGG AGATCGCCTA CTATCGCCTT ACCTCGCCGT CGCTCGCCTC
 GGGTCGCCTG TGTTCGCCTT AAAAGGCCTT ACGAGGCCAT TATAGGCCTA
 GCTAGGCCCA TGACGGCCGG CTTCGGCCGG CAAGGGCCTC AACGGGCCTA
 35 GCCGGGCCAG CCGGGGCCAC ATATGGCCCG TTATGGCCAA TAGTGGCCGC

GTGTGGCCCA CGTTGGCCAA ATTTGGCCAT GGAATGCCAC GTCATGCCAT
TGGATGCCCC TTA CTGCCTT TGCCTGCCTC ACTCTGCCGC CGTCTGCCTT
ATTCTGCCTA GTAGTGCCAA CACGTGCCGG GAGGTGCCGC ATGGTGCCAG
TTGGTGCCCA CCATTGCCCC ACCTTGCCCA TCCTTGCCAG AGAAATCCAA
5 TGAAATCCCG CTAAATCCAT CGCAATCCAC GGAATCCAG TGTAATCCAT
CTTAATCCCG GCTCATCCGA GTTCATCCTC GAAGATCCGG ATAGATCCAC
AGCGATCCAT CTCGATCCAA AAGGATCCGA TACTATCCTC GCCTATCCCC
CCGTATCCCG CATTATCCTT CGACCTCCAA ATACCTCCAT AGCCCTCCAC
ATTCTCCCG TTTCCTCCAA TTGGCTCCTG GTTGCTCCTT CTATCTCCGG
10 GACTCTCCAA ACCTCTCCTG GGGTCTCCGT TGTCTCCGG AAAAGTCCGG
GTAAGTCCAC GGCAGTCCAT GAGAGTCCGA TTGAGTCCCC TGACGTCTT
TCCCGTCCGA TTCCGTCTC AGTCGTCTA CTTCGTCTT GGAGGTCCAA
AGGGGTCCAG TGGGGTCCCA CCCTGTCCCT GATTGTCTC TCTGTCCCC
CAAATCCTA AACATCCTC CCGATTCCAA TTA CTTCGG AAGCTTCCCC
15 CGTCTTCCGG CACGTTCTT TATGTTCTC GCTGTTCCCC AGATTTCGG
TGATTTCBA GGGTTTCCCA AGTTTTCCAT CTTTTTCCAA TAAAAAGCTC
GCAAAAGCCC TCGAAAGCAC CCACAAGCTG GACCAAGCAC ATCCAAGCGG
CAGCAAGCAG CATGAAGCTA CGATAAGCAC AGCTAAGCAA TGCTAAGCCG
CTCTAAGCAT AAGTAAGCGT TTTTAAGCAC GAAACAGCAA ACAACAGCTG
20 CTCACAGCGG AAGACAGCAG TAGACAGCCA GCGACAGCTA GTGACAGCGC
TACCCAGCTA GCCCCAGCCA AGAGCAGCGG GACGCAGCAT GGGGCAGCGA
GCTTCAGCGC GTTTCAGCTA CCAAGAGCCT CCTAGAGCAG CACCGAGCAT
CGGCGAGCGA ATGCGAGCGT TATCGAGCAC ACAGGAGCCC CCCGGAGCCA
GGATGAGCAT AACTGAGCGG GTCTGAGCAC TGGTGAGCCT CTGTGAGCAG
25 GGTTGAGCCG AGCATAGCCG TGCATAGCAA TAGATAGCGT GATATAGCGG
ATTATAGCAC GGACTAGCTA GCCCTAGCGT AGGCTAGCTG TGAGTAGCAT
CTAGTAGCCG TTCGTAGCAC GGGGTAGCCT AGTGTAGCAA TGTGTAGCCG
CTTGTAGCAT AAATTAGCTC CACTTAGCTA ACGTTAGCAC AAAAACGCTG
GCAAACGCAA GAGAACGCTA ACGAACGCAG TCGAACGCCA TGACACGCGT
30 GACCACGCCA TCCCACGCTA TTCCACGCGC AGTCACGCGA CTTACGCGT
GCCGACGCAT CGATACGCCA ATATACGCCT AGCTACGCCC GATTACGCGC
ATTTACGCAG TTTTACGCCA GAAACCGCCC TCAACCGCTC TTAACCGCGA
TAGACCGCAC CGTACCGCGA ATTACCGCGT CAACCGCTG GCCCCGCGC
CCGCCCCGAG TGAGCCGCGC TTCGCCGCGT CCTGCCGCTA CTTGCCGCGC
35 GGCTCCGCTG ACGTCCGCGT CGGTCCGCTC ACCAGCGCAT AAACGCGCCT

AATCGCGCAG TATCGCGCCA GCTCGCGCTA GTTCGCGCGC GAAGGCGCTG
ACAGGCGCAA TCAGGCGCCG CCCGGCGCAC AAGGGCGCTA GCGGGCGCAG
TCTGGCGCAT TACTGCGCGC GTCTGCGCCA CATTGCGCGT CGAATCGCAG
TTAATCGCCT TGCATCGCCC GTGATCGCAT CGTATCGCCT ATTATCGCCA
5 TTTATCGCAG TGGCTCGCTC GACGTCGCGC ATCGTCGCAG TTCGTCGCCA
CAGGTCGCGG AGTGTCGCCC TAATTCGCTG GCATTGCGCG ACGTTCGCCA
TCGTTGCGCAG GCTTTCGCAT TTGAAGGCTG GTTAAGGCTT CTACAGGCAC
CGCCAGGCAT ATCCAGGCAA TTCCAGGCCG CAGCAGGCGA TGTCAGGCAC
GAATAGGCAT AGCTAGGCGG TAGTAGGCCT GCGTAGGCTT GATTAGGCCG
10 TCTTAGGCTG GAAACGGCGG ATAACGGCAC AGCACGGCAT CTCACGGCAA
AAGACGGCGA GTCCCGGCTA AGAGCGGCAA TGAGCGGCCG CTAGCGGCAT
CGCGCGGCAC GGGGCGGCAG TGTGCGGCAT CTTGCGGCCG AATTCGGCTC
TGAAGGGCTC TTCAGGGCTT CCTAGGGCGA CTTAGGGCTC GTACGGGCAT
GGCCGGGCAC CGGCGGGCAG TTGCGGGCCT GTTCGGGCCG TCAGGGGCGC
15 TTAGGGGCTA CGTGGGGCTA ATTGGGGCTT CAATGGGCAC AACTGGGCAA
TACTGGGCCG GCCTGGGCTG CCGTGGGCTC AGGTGGGCGT CTGTGGGCGA
TGCATGGCGG AAGATGGCCT GATATGGCAA ACTATGGCTG AACCTGGCTT
GACGTGGCCG TCCGTGGCTG CAGGTGGCCC AGTGTGGCGG GCATTGGCGC
GTATTGGCTA ATGTTGGCTG TTGAATGCGT AATAATGCAC GTTAATGCGG
20 CCTCATGCCA CAAGATGCCG TACGATGCAC CATGATGCAT ACATATGCGA
ATATATGCTC CCCTATGCGC AGCTATGCTT CTCTATGCTA GCGTATGCGG
TCTTATGCGT GAAACTGCTT TCAACTGCCT GCGACTGCAT ACTACTGCCA
TCTACTGCAG TACCCTGCAT TGGCCTGCGA CATCCTGCAC GACGCTGCTA
ACCGCTGCAG TCCGCTGCCA GTATCTGCCG CGGTCTGCCT ATGTCTGCCA
25 TTGTCTGCAG GTTCTGCAT TTCAGTGCGG CAGAGTGCCA AAACGTGCTC
CACCGTGCTA ACGCGTGAC GATGGTGAC ATTGGTGCGG GGATGTGCTA
GCCTGTGCGT AGGTGTGCTG TCAATTGCGA TTAATTGCTC TGCATTGCTT
ACTATTGCGT CGTATTGCTC GGACTTGCAT AACCTTGCGG GTCCTTGAC
TGGCTTGCTT CTGCTTGCGG GGTCTTGCCG TGAGTTGCTA TCCGTTGCGT
30 CCTGTTGCGC AGTGTGCTT CTTGTTGCTA CACTTTGCAT CGGTTTGCGA
ATGTTTGCGT TATTTTGAC GAAAAATCGC ATAAAATCAG TAAAATCCA
GTGAAATCAA CGTAAATCCA ATTAAATCCT TGAGAATCCC CGCGAATCAG
TTCGAATCCT GGGGAATCAC CTTGAATCCC ACGTAATCCT AATTAATCTG
GCTTAATCAA TCGACATCGA TTGACATCTC TGACCATCCT CTACCATCAG
35 GACCCATCGA TTCCCATCCC GGGCCATCAT AGTCCATCCA TGTCCATCAG

CTTCCATCCT TGGGCATCTA CGATCATCGA ATATCATCGT CCCTCATCTT
 AGTCATCGC GATTCATCCC TCTTCATCTC TTTTCATCGA CAAAGATCGT
 AGGAGATCAC ATTCGATCTG AAAGGATCGC GTAGGATCAG CACGGATCGA
 CGGGGATCAT ATGGGATCAA TTGGGATCCG GTTGGATCCT ACCTGATCCG
 5 TCCTGATCAA GCAATATCAT GAGATATCTT TCGATATCCT TATATATCTG
 GCTATATCCG TGAATATCGA GACCTATCCT TCCCTATCTT CCTCTATCTC
 AGTCTATCGT CTTCTATCGA AACGTATCTG GCCGTATCAA CGATTATCCT
 ATATTATCCA TTATTATCAG CTCTTATCCC GTGTTATCCG CGTTTATCAG
 TTTTATCCT TTAAACTCAC TGCAACTCAT CTCAACTCCG TAGAACTCGA
 10 GTGAACTCCC CGTAACTCAC GGACACTCTT GCCCACTCGA GTCCACTCTC
 CTGCACTCTG AGAGACTCCG TGAGACTCAA GGGGACTCCA AGTGACTCAT
 CTTGACTCAA CACTACTCTT TATTACTCTC GCTTACTCCC GCAACCTCGC
 GTAACCTCTA ATGACCTCTG ATCCCCTCCG TTCCCCTCAA AGTCCCTCAC
 GGAGCCTCTC GTCGCCTCTT TGCTCCTCGG AAGTCCTCCT GATTCCTCAA
 15 ACTTCCTCTG GGAAGCTCCG AGGAGCTCCA TGGAGCTCAG CTGAGCTCCT
 GGTAGCTCAT ACACGCTCGT CGACGCTCTC AGCCGCTCTA CTCCGCTCTT
 TCTCGCTCGA TTTCGCTCTC TAAGGCTCGG ATGGGCTCCC CCATGCTCAG
 GACTGCTCTC TCCTGCTCCC CAGTGCTCTG CTTGCTCCT AATATCTCTC
 CGCCTCTCGG GGGCTCTCGC TACGTCTCTC GCCGTCTCCC CCGGTCTCCG
 20 CATGTCTCTT GAATTCTCGG ATATTCTCAC AGCTTCTCAT CTCTTCTCAA
 AAGTTCTCGA ACAAAGTCTC CGAAAGTCGT ATAAAGTCGA CCCAAGTCTA
 CTCAAGTCGC AAGAAGTCAC GTGAAGTCGG TTAAAGTCGT CCGCAGTCCA
 CCAGAGTCTT AGAGAGTCGC CGCGAGTCGA ATCGAGTCGT CAGGAGTCAT
 TCGTAGTCGT GCTTAGTCGG TAAACGTCTG GCAACGTCCG ACGACGTCCA
 25 TCGACGTCAG GCTACGTCAT CCACCGTCTC AGACCGTCGT CTACCGTCGA
 GACCCGTCAG ACCCCGTCTA ATCCCGTCGC CAGCCGTCAC TGTCCGTCGA
 CGATCGTCAG TTATCGTCCT TGCTCGTCCC GTGTCGTCAT CGTTCGTCCT
 ATTTCTGTTA TTTCGTCAG GGAAGGTCGC GTCAGGTCGT CATAGGTCCA
 ACTCGGTCAC TAAGGGTCCC GCAGGGTCTC GTAGGGTCGA CACGGGTCAG
 30 GAGGGGTCAC ATGGGGTCGG TCCTGGTCGG AGTTGGTCTG GGCATGTCTG
 ACGATGTCGT CGGATGTCTC AGACTGTCCA TGAATGTCAG CTACTGTCCT
 CGCCTGTCCC GGGCTGTCCG TGTCTGTCCT CTTCTGTCAG GCCGTGTCCG
 CCGGTGTGCG AGGGTGTCTT CTGGTGTCTA GAATTGTCCC TCATTGTCTC
 TTATTGTCGA TAGTTGTCAC CGTTTGTGCG ATTTTGTCTG CGAAATTCTG
 35 GTGAATTCTT TTAAATTCTG GGACATTCCC GTCCATTCTT TGGCATTAC

CATCATTCGA CCAGATTCGG ATCGATTCTG AAATATTCCA TAATATTCAG
CACTATTCCC GAGTATTCCG TCGTATTCTG TATTATTCCT GCTTATTCTT
TAAACTTCGT GGCAC TTC AATACTTCGA AGACCTTCTG GGGCCTTCTA
GGAGCTTCCT TACGCTTCGA GTCGCTTCCC TGGGCTTCAT CTGGCTTCCG
5 GGTGCTTCAG ACATCTTCAC CCCTCTTCAA AAGTCTTCTC GTCAGTTCTG
CCGAGTTCGA CTGAGTTCTC CGACGTTCCCT ATACGTTCCA TTACGTTCAG
CTCCGTTCCC GTGCGTTCCG CGTCGTTCAG TTTCGTTCCCT CGGGGTTCTA
ATGGGTTCTT TGATGTTCTG GACTGTTCCCT TCCTGTTCTT CCTTGTTCTC
AGTTGTTCTG CTTTGTTCTG AAAATTTCTG TAAATTTCCA GCAATTTCTA
10 GTAATTTCTG GGCATTTCTG GAGATTTCAA ACGATTTCTG AATATTTCTT
ACCCTTTCCG TCCCTTTCAA GGAGTTTCTG TACGTTTCCCT GCCGTTTCTT
AGGGTTTCCG CATGTTTCCC ATTTTTTCTG TGAAAAAGGA GACAAAAGCT
TCCAAAAGTT CCTAAAAGTC AGTAAAAGGT CTAAAAGGA GCACAAAGAT
GAGCAAAGTT TCGCAAAGCT TATCAAAGTG GCTCAAAGCG GAAGAAAGCA
15 TCAGAAAGTA TTAGAAAGGC TCGAAAGGT AAGGAAAGCG TAGGAAAGAA
ACTGAAAGTT CGTGAAAGGC GTCTAAAGTG CCGTAAAGGA CTGTAAAGTC
TTACCAAGGT TGCCCAAGGC ACTCCAAGTC CGTCCAAGGT ATTCCAAGGA
GCAGCAAGAC CACGCAAGTG GAGGCAAGTC TCGGCAAGCC AGATCAAGCT
CTATCAAGCA ATCTCAAGCC AGTTCAAGAG TGTTCAGCA AAAAGAAGCA
20 TAAAGAAGAG CACAGAAGCC GAGAGAAGCG TCGAGAAGTG TATAGAAGCT
GCTAGAAGTT CCACGAAGAC GACCGAAGTG ACCCGAAGAA TCCCGAAGCG
CAGCGAAGTC AACGGAAGCT TGGGGAAGGG GGTGGAAGGT CGATGAAGTG
GTGTGAAGTT TTTTGAAGTG GAACTAAGGC ATACTAAGAG TTAATAAGCA
GTGCTAAGAA CGTCTAAGCA ATTCTAAGCT TTGGTAAGTT GTTGTAAGTG
25 AGATTAAGGA CTATTAAGGT ACCTTAAGTT CGCTTAAGGC GGGTTAAGGG
TGTTTAAGGT CGCAACAGGG GGAACAGGC AATCACAGTC GAAGACAGAC
ATAGACAGGG CCCGACAGTG TAGGACAGCC GCGGACAGTC GTGGACAGGA
GGTTACAGTA TACACCAGTT GCCACCAGCT CATAACAGTC GAACCCAGAT
AGCCCCAGGG TAGCCCAGCT GCGCCCAGTT GATCCCAGCG TCTCCCAGTG
30 TAAGCCAGTA GCAGCCAGCA ACGGCCAGCG TCGGCCAGAA AATGCCAGTT
CTATCCAGAC CGCTCCAGAT ATCTCCAGAA TTCTCCAGCG CAGTCCAGGA
TGTTCCAGAC AAAAGCAGAC GTAAGCAGGG CACAGCAGAA ACGAGCAGTC
CGGAGCAGGT ATGAGCAGGA CCACGCAGCA ACCCGCAGCC CAAGGCAGAT
CCGGGCAGTT AGGGGCAGGC CATGGCAGCG TCATGCAGGT TGCTGCAGTA
35 ACTTGCAGGA ATTTGCAGTC GGAATCAGTT GCCATCAGGA GTCATCAGTC

CTGATCAGTG TTA CTCAGAC TGCCTCAGAT CTCCTCAGCG TAGCTCAGGA
GTGCTCAGCC CGTCTCAGAC GCAGTCAGGT GCGGTCAGTA CAGTTCAGCT
CCTTTCAGTG AGAAAGAGCA TGAAAGAGAG CTAAAGAGCT CGCAAGAGCC
GGGAAGAGCG TGTAAGAGCT CTTAAGAGAG GGCCAGAGTG ACGCAGAGGT
5 CGGCAGAGTC ATAGAGAGCC AGCGAGAGCT CTCGAGAGCA TAGGAGAGGG
GATGAGAGGT CAATAGAGTG GCCTAGAGAC CCGTAGAGAG CCGACGAGGT
AGGACGAGTC CGACCGAGCA ATACCGAGCT AGCCCGAGCC GATCCGAGGC
ATTCCGAGAG TTTCCGAGCA GTAGCGAGTG ACGGCGAGGC CGGGCGAGTT
ATGGCGAGTA TGATCGAGGT GACTCGAGCA TCCTCGAGTA TTCTCGAGGC
10 AGTTCGAGGA CTTTCGAGGT TAAAGGAGGA GTAAGGAGCC GGCAGGAGCT
TTGAGGAGAC AATAGGAGGT CTACGGAGTG ACCCGGAGGG GGGCGGAGTT
TGTCGGAGTG GGAGGGAGCA TACGGGAGGT AGGGGGAGCG TGGGGGAGAA
CATGGGAGGC CCCTGGAGAT TCTTGGAGAC CCGATGAGCA AACTGAGTC
CGACTGAGGT ATACTGAGGA CCCCTGAGTA CTCCTGAGGC AAGCTGAGAC
15 GTGCTGAGGG TTTCTGAGGT GCTGTGAGAC AGATTGAGAG TGATTGAGCA
GACTTGAGGT GGGTTGAGAA AGTTTGAGCT CTTTGTAGCA ACCAATAGCG
TCCAATAGAA AAACATAGAG TAACATAGCA GCACATAGTA GTACATAGGC
GGCCATAGGT GAGCATAGAA ACGCATAGTG AATCATAGCT TCAGATAGAT
TAGGATAGTT GCGGATAGCT GATGATAGTG ACTGATAGAA TCTGATAGCG
20 CAATATAGGT AGGTATAGAC CAAACTAGAG TACACTAGCC GCCACTAGTC
GTCACTAGGA CCGACTAGTG CATACTAGCT TCACCTAGAC CCCCCTAGCG
TAGCCTAGTC GCGCCTAGCC GTAGCTAGGT GCGGCTAGGC CGGGCTAGGG
AATGCTAGCC TGATCTAGTG CTTTCTAGTG ACGAGTAGCT AATAGTAGTG
GCTAGTAGAA AGACGTAGGA CTACGTAGGT ACCCGTAGTT CGCCGTAGGC
25 GGGCGTAGGG TGTCGTAGGT TACGGTAGTG GCCGGTAGCG CCGGGTAGCC
GAATGTAGGC ATATGTAGAG TTATGTAGCA GTGTGTAGAA CGTTGTAGCA
ATTTGTAGCT GGAATTAGCC GTCATTAGCT TGGATTAGAC CATATTAGGA
CGACTTAGTG GTGCTTAGTT TTTCTTAGTG TAAGTTAGGC GTAGTTAGCA
ATGGTTAGCG TTGGTTAGAA CCATTTAGAC GACTTTAGTG ACCTTTAGAA
30 TCCTTTAGCG CAGTTTAGTC CGAAAACGGG GTGAAACGGT TTAAACGGG
CATCAACGTA CCAGAACGTG GACGAACGAC ATCGAACGGG CAGGAACGAG
TCGTAACGGG GCTTAACGGT TAAACACGTT GCAACACGCT TCGACACGAT
AATACACGTA GCTACACGAG AGACCACGGG GACCCACGAT GGGCCACGGA
TACGCACGTA GCCGCACGCA CGATCACGAT ATATCACGAA TTATCACGCG
35 CTCTCACGAC AAGTCACGGC GTGTCACGAG CGTTCACGCG TTTTCACGAT

CAAAGACGAA AACAGACGAC GTCAGACGGG CCGAGACGTA CTGAGACGGC
ACACGACGCC CCCCAGCGCA CACGGACGAT CGGGGACGGA ATGGGACGGT
TATGGACGAC TGATGACGTA TCCTGACGGT CCTTGACGGC AGTTGACGTT
CTTTGACGTA GCAATACGGA GTAATACGTC GGCATACGTT ACGATACGGG
5 TGACTACGAT CTACTACGCG TTCCTACGAC GGGCTACGCT AGTCTACGAA
TGTCTACGCG CTTCTACGAT GGAGTACGTA GCCGTACGGT AGGGTACGTG
GATTTACGAC ATTTTACGGG GAAAACCGCT TCAAACCGTT TGCAACCGGA
TAGAACCGAT GATAACCGAG ACTAACCGTA ATTAACCGGC GCCCACCGAT
TGAGACCGGT GACGACCGCA TCCGACCGTA TTCGACCGGC AGTGACCGGA
10 CTTGACCGGT GTATACCGTG ACGTACCGGC CGGTACCGTT ATGTACCGTA
TGACCCCGGC TTCCCCCGGT CCTCCCCGTA CTCCCCCGGC CAAGCCCGTG
GCCGCCGAC CCGGCCGAG ATATCCCGCC AGCTCCCGCT CTCTCCCGCA
TAGTCCCGGG GATTCCCGGT CAAAGCCGCC AACAGCCGCA TACAGCCGAG
GGTAGCCGGA GAACGCCGTG ACACGCCGAA TCACGCCGCG CCCCGCCGAC
15 AAGCGCCGTA GCGCGCCGAG TCTCGCCGAT AAAGGCCGCT AATGGCCGAG
TATGGCCGCA GCTGGCCGTA GTTGGCCGGC CGCTGCCGTG GGGTGCCGTC
TCGATCCGGC TTGATCCGTA GACCTCCGGC ATCCTCCGAG TTCCTCCGCA
CAGCTCCGGG AGTCTCCGCC TGGGTCCGTC ACATTCCGTT CGATTCCGGC
AGCTTCCGGA CTCTTCCGGT AAGTTCCGAT GATTTCCGCA TCTTTCCGTA
20 TTTTCCCGC CGAAAGCGAA ATAAAGCGAT AGCAAGCGAC ATTAAGCGCG
TTTAAGCGAA CTAGAGCGAC CGCGAGCGAT ATCGAGCGAA TTCGAGCGCG
CAGGAGCGGA TGTGAGCGAC TAATAGCGTA GCATAGCGCA ACGTAGCGCG
TCGTAGCGAA AATTAGCGTT GCTACGCGGA GTTACGCGTC AGACCGCGAA
TGACCGCGCG CTACCGCGAT CGCCCGCGAC GGGCCGCGAG TGTCCGCGAT
25 CTCCGCGCG GGTGCGCGTA GAATCGCGAC ATATCGCGGG CCCTCGCGTG
TAGTCGCGCC GCGTCGCGTC GTGTCGCGGA CAAAGGCGGG GTCAGGCGAA
TCACGGCGGC TTACGGCGTA CGTCGGCGTA ATTCGGCGTT GTAGGGCGAT
GGCGGGCGAC CGGGGGCGAG TTGGGGCGCT GTTGGGCGCG ACCTGGCGCT
AAAATGCGTA GCAATGCGAG CACATGCGTC GAGATGCGTG ACGATGCGAA
30 TCGATGCGCG TATATGCGTT GCTATGCGCT GACCTGCGCG TCCCTGCGTG
CAGCTGCGCC AGTCTGCGGG AACGTGCGTT CGATTGCGCG TTATTGCGAT
TGCTTGCGAC GTGTTGCGCT CGTTTGCGAT ATTTTGCGAA TTTTGCGCG
GAAAATCGTC TCAAATCGCC CCAATCGAG GCGAATCGAC CAACATCGCG
TACCATCGAC CATCATCGAT CCTGATCGCA GGCTATCGCG CGGTATCGCC
35 AATTATCGGG GTTTATCGAC AAAACTCGCG TAAACTCGAA GAGACTCGCA

TCGACTCGTA TTGACTCGGC AATACTCGAT GACCCTCGTA ACCCCTCGAG
TCCCCTCGCA TACGCTCGAT TGGGCTCGGA CATGCTCGAC CGATCTCGTA
ATATCTCGTT CCCTCTCGGT AGCTCTCGTC TCTTCTCGGC TTTTCTCGTA
CAAAGTCGTT CCGAGTCGAT GATCGTCGAC ATTCGTCGGG AAAGGTCGTC
5 CACGGTCGTA ACGGGTCGAC TGATGTCGAT CTATGTCGCG TTCTGTCGAC
GGGTGTCGCT AGTTGTCGAA TGTTGTCGCG CTTTGTGCGAT GGCATTGCGAA
GAGATTCGGT TATATTCGGG TGAATTCGTA TCCCTTCGGT CCTCTTCGGC
AGTCTTCGTT CTTCTTCGTA GGAGTTCGAT AACGTTTCGGG GTCGTTTCGAC
TGGGTTCGCT CTGGTTCGAG GGTGTTTCGCG ACATTTTCGCC CCCTTTTCGCA
10 GGAAAAGGCG AGGAAAGGCA TGGAAAGGAG CTGAAAGGCT GGTAAGGAT
ACACAAGGGT CGACAAGGTC AGCCAAGGTA CTCCAAGGTT TCTCAAGGGA
TTTCAAGGTC TAAGAAGGGG ATGGAAGGCC CCATAAGGAG GACTAAGGTC
TCCTAAGGCC CAGTAAGGTG CCTTAAGGCT CCAACAGGGT AGAACAGGTC
CGCACAGGTA ATCACAGGTT CGGCCAGGCA ATGCCAGGCT GTTCCAGGAA
15 ACAGCAGGGC CGAGCAGGTT ATAGCAGGTA CCCGCAGGGA CTCGCAGGTC
GTGGCAGGTG TTTGCAGGTT AACTCAGGCG TACTCAGGAA TGGTCAGGGT
GGTTCAGGGG TTAAGAGGAC TGCAGAGGAT CTCAGAGGCG TAGAGAGGGA
GTGAGAGGCC CGTAGAGGAC GGACGAGGTT GCCCGAGGGA GTCCGAGGTC
CTGCGAGGTG AGAGGAGGCG TGAGGAGGAA GGGGGAGGCA AGTGGAGGAT
20 CTTGGAGGAA CACTGAGGTT TATTGAGGTC GCTTGAGGCC CCAATAGGCA
ACCATAGGCC AAAGTAGGAC GTACTAGGGG CACCTAGGAA ACGTAGGTC
CGGCTAGGGT ATGCTAGGGA CCCGTAGGCT GATGTAGGTC TCTGTAGGCC
GGATTAGGAA AGGTTAGGAG TGGTTAGGCA CAAAACGGGT AGGAACGGAC
ATTCACGGTG AAAGACGGGC GTAGACGGAG CACGACGGGA CGGGACGGAT
25 ATGGACGGAA TTGGACGGCG GTTGACGGCT ACCTACGGCG TCCTACGGAA
TGAACCGGTG CTTACCGGTG AAACCCGGGT GGCCCCGGAG CGGCCCGGAC
TATCCCGGGA GTTCCCGGCC TCAGCCGGGG CAATCCGGAG TACTCCGGCC
GCCTCCGGTC GTCTCCGGGA CCGTCCGGTG CATTCCGGCT GAAAGCGGGC
ATAAGCGGAG TTAAGCGGCA GTGAGCGGAA CGTAGCGGCA ATTAGCGGCT
30 TGAGGCGGCC CGCGGCGGAG TTCGGCGGCT GGGGGCGGAC CTTGGCGGCC
ACGTGCGGCT AATTGCGGTG GCTTGCGGAA CCAATCGGAC GACATCGGTG
ACCATCGGAA TCCATCGGCG CAGATCGGTC AAATCGGCA TAACTCGGAG
CACCTCGGCC GAGCTCGGCG TCGCTCGGTG TATCTCGGCT GCTCTCGGTT
ACAGTCGGAT AAGGTCGGTT ACTGTGCGCG TCTGTCGGAA GGATTCCGCC
35 GTCTTCGGCT TGGTTCGGAC CATTTCGGGA TACAAGGGAT TGGAAGGGGA

CATAAGGGAC GAACAGGGTT TCACAGGGCT GCGCAGGGAT ACTCAGGGCA
TCTCAGGGAG AAAGAGGGCG TAAGAGGGAA GAGGAGGGCA TCGGAGGGTA
TTGGAGGGGC AATGAGGGAT CCATAGGGGA CTATAGGGTC ACCTAGGGGC
CGCTAGGGTT ATCTAGGGTA TGTTAGGGTC CCTACGGGCA TTGCCGGGGT
5 AATCCGGGAC GTTCCGGGGG GAAGCGGGTC TCAGCGGGCC CCCGCGGGAG
GCGGCGGGAC TACTCGGGGG AGGTCGGGCT CTGTCTGGGCA GGTTCGGGAA
GAAAGGGGCG TCAAGGGGTG AAGAGGGGCA TAGAGGGGAG GATAGGGGAT
CAACGGGGTC AACCGGGGTA GCGCGGGGAG CCGCGGGGAC TGAGGGGGGG
CTTGGGGGGG GTATGGGGTT GGCTGGGGTC CGGTGGGGTG TCCATGGGGC
10 TTCATGGGTA CCTATGGGGT AGTATGGGTC GTACTGGGAA CACCTGGGGG
GAGCTGGGGC ATGCTGGGAG TTGCTGGGCA TTAGTGGGTT TGCCTGGGTC
ACTGTGGGGC CGTGTGGGTT ATTGTGGGTA GGATTGGGGG AACTTGGGAT
AGGTTGGGGA CTGTTGGGGT GGTAATGGTA GAACATGGGG ATACATGGAC
AGCCATGGAT CTCCATGGAA AAGCATGGGA GCTGATGGGA GTTGATGGTC
15 CGCTATGGGG GGGTATGGGC CTAAGTGGAC CGCACTGGAT ATCACTGGAA
TTCATGGCG CAGACTGGGA TGTACTGGAC TTGCCTGGTG GTTCCTGGTT
CGAGCTGGAA ATAGCTGGAT AGCGCTGGAC ATTGCTGGCG TTTGCTGGAA
TACTCTGGTT GCCTCTGGCT CATTCTGGTC TCAAGTGGGT TGCAGTGGTA
ACTAGTGGGA ATTAGTGGTC GGACGTGGAA AGGCGTGGAG TGGCGTGGCA
20 TGAGGTGGTT TCCGGTGGGA TTCGGTGGTC AGTGGTGGTA CTTGGTGGTT
AAATGTGGAC GTATGTGGGG CACTGTGGAA ACGTGTGGTC CGGTGTGGGT
ATGTGTGGGA CAGATTGGCT CCTATTGGTG CACCTTGGTT TATCTTGGTC
GCTCTTGGCC TTAGTTGGGG AAGGTTGGCC CGTGTGGGGG GGATTGGTT
GCCTTTGGGA GTCTTTGGTC CTGTTTGGTG ACCAAATGAT AAACAATGCT
25 AATCAATGAG TATCAATGCA GCTCAATGTA GTTCAATGGC GAAGAATGTG
ACAGAATGAA TCAGAATGCG CCCGAATGAC AAGGAATGTA GCGGAATGAG
TCTGAATGAT TACTAATGGC GTCTAATGCA CATTAAATGGT CAAACATGCT
CATAATGAG CCCCCATGAT TCTCCATGAC AAAGCATGCC CACGCATGCA
TTGGCATGGG GTTGCATGGT CTATCATGTG ACCTCATGGG GGGTCATGTT
30 TGTTTCATGTG AAAAGATGTG GCAAGATGAA GAGAGATGTA ACGAGATGAG
TCGAGATGCA TGACGATGGT GACCGATGCA TCCCGATGTA TTCCGATGGC
AGTCGATGGA CTTGATGGT GCCGGATGAT CGATGATGCA ATATGATGCT
AGCTGATGCC GATTGATGGC ATTTGATGAG TTTTGATGCA CAAATATGGA
AACATATGGC GTCATATGAG CTGATATGAC GGTATATGCC TTACTATGTG
35 CGTCTATGTG CACGTATGGT CGGGTATGAA ATGGTATGAT TATGTATGGC

GTTGTATGCA TCCTTATGAT CCTTTATGAC CCAAACCTGCT CCTAACTGAG
CACCACCTGAT CGGCACTGGA ATGCACTGGT TATCACTGAC ACAGACTGCC
CCCGACTGCA GGATACTGAT AACTACTGGG GTCTACTGAC TGGTACTGCT
CTGTACTGAG GGTTACTGCG GGAACCTGGG AACACCTGAT AGGACCTGGA
5 CTGACCTGGT ACACCCTGCT GATCCCTGTA ACTCCCTGAG TCTCCCTGCA
AAAGCCTGAA TAAGCCTGCG GCAGCCTGTG CACGCCTGAC GAGGCCTGAG
ACGGCCTGTA ATGGCCTGGC TATGCCTGAT TCCTCCTGGC TTCTCCTGTA
CCTTCCTGGT AGTTCCTGTC TAAAGCTGTC GCAAGCTGCC TCGAGCTGAC
CCACGCTGTG GACCGCTGAC ATCCGCTGGG CAGCGCTGAG CATGGCTGTA
10 CGATGCTGAC AGCTGCTGAA TGCTGCTGCG CTCTGCTGAT AAGTGCTGGT
TTTTGCTGAC TACATCTGGG AGGATCTGCT CTGATCTGCA GGTATCTGAA
ACACTCTGGA ATACTCTGTC CCCCTCTGGC AGCCTCTGTT CTCCTCTGTA
GCGCTCTGGG TCTCTCTGGT GCGCTCTGCG CGGGTCTGCC AATGTCTGGG
GTTGTCTGAC CCTTTCTGCA AGAAAGTGTG GGGAAAGTGTA TAACAGTGGT
15 GGCCAGTGCA AATCAGTGGA ACAGAGTGGG CTCGAGTGTG GCGGAGTGGA
GTGGAGTGTC CAATAGTGCA AACTAGTGCC GGTTAGTGGC GGAACGTGCC
GTCACGTGCT TGGACGTGAC CACACGTGGA CGACCGTGTG GTGCCGTGTT
TTTCCGTGTG TAAGCGTGGC GTAGCGTGCA ATGGCGTGCG TTGGCGTGAA
CCATCGTGAC GACTCGTGTG ACCTCGTGAA TCCTCGTGCG CAGTCGTGTC
20 GCAAGGTGGG ACGAGGTGGA ATGAGGTGTC AGACGGTGCT CTACGGTGCA
ATCCGGTGCC AGTCGGTGAG TGTCGGTGCA GGAGGGTGTG AGGGGGTGTA
CTGGGGTGTT TTATGGTGGT TGCTGGTGGC ACTTGGTGTC CGTTGGTGGT
ATTTGGTGGA CAAATGTGAG TACATGTGCC GCCATGTGTC GTCATGTGGA
CCGATGTGTG CATATGTGCT TCACTGTGAC CCCCTGTGCG TAGCTGTGTC
25 GCGCTGTGCC GTAGTGTGGT GGCGTGTGGC CGGGTGTGGG AATGTGTGCC
TGATTGTGTG CTTTTGTGTG CCAAATTGTC AGAAATTGGT CTAAATTGGA
GACAATTGAG ACCAATTGTA ATCAATTGGC CAGAATTGAC TGTAATTGGA
TAACATTGTG GCACATTGCG ACGCATTGCA TCGCATTGAG GCTCATTGAT
ACAGATTGTT CGAGATTGGC AGCGATTGGA CTCGATTGGT AAGGATTGAT
30 GATGATTGCA TCTGATTGTA TTTGATTGGC TGGTATTGTC CCGACTTGCA
ACACCTTGTC CGACCTTGGT ATACCTTGGA CCCCTTGTA CTCCCTTGGC
AAGCCTTGAC GTGCCTTGGG TTTCTTGGT GCTGCTTGAC AGATCTTGAG
TGATCTTGCA GACTCTTGGT GGGTCTTGAA AGTTCTTGCT CTTTCTTGCA
TAAAGTTGCT GCAAGTTGTT GGCAGTTGGA GAGAGTTGAT AATAGTTGCA
35 TATAGTTGAG TCCCGTTGAT CCTCGTTGAC GGAGGTTGGT AACGGTTGAG

TACGGTTGCA GCCGGTTGTA GTCGGTTGGC CTGGGTTGGG TTATGTTGTG
CGTTGTTGTG CCGATTTGGT AGGATTTGTC CGACTTTGCA ATACTTTGCT
AGCCTTTGCC GATCTTTGGC ATTCTTTGAG TTTCTTTGCA GTAGTTTGTG
ACGGTTTGGC CGGGTTTGTG ATGGTTTGTG TGATTTTGGT GACTTTTGTG
5 TCCTTTTGTG TTCTTTTGGC AGTTTTTGGG CTTTTTTGGT GATAAAATAC
ATTAAAATGG CAACAAATTT CCGCAAATAT TTCGAAATGG CAGGAAATCA
GCATAAATGA GTATAAATTC GGCTAAATTT ACGTAAATGG AAAACAATTT
ACGACAATAT TATACAATTA GCTACAATCA TGACCAATGG CTTCCAATGG
CAAGCAATTC AACGCAATTA GCCGCAATAG CCGGCAATAC ATATCAATCG
10 TTATCAATAA TAGTCAATGC GTGTCAATCA CGTTCAATAA ATTTCAATAT
CAAAGAATCG TACAGAATAC CATAGAATAT GAACGAATTC TCACGAATCC
CCCCGAATAG GCGCGAATAC TTGGGAATGT AATGGAATAC GTTGGGAATGG
AGATGAATTA CTATGAATTT ACCTGAATGT CGCTGAATTC GGGTGAATTG
TGTTGAATTT TCGATAATGG GCTATAATGT AGACTAATAT CTAATAATAA
15 GACCTAATGG ATCCTAATAC CAGCTAATGC AGTCTAATCG TGTCTAATAA
TGGGTAATTG GGTGTAATTT CGATTAATGG GTGTTAATGT TTTTAAATGG
ACAAACATTT CGAAACATGC AGCAACATGA CTCAACATGT AAGAACATAT
GATAACATCA TCTAACATTA TTAAACATGC CCAGACATTC AGAGACATGT
CTAGACATGA GACGACATAG ACCGACATTA ATCGACATGC CAGGACATAC
20 TGTGACATGA TCGTACATGC TTGTACATTA GCTACCATAC CCACCCATTT
AGACCCATGC CGCCCCATGA ATCCCCATGT CAGCCCATAT CCGGCCATCA
GAATCCATGA TTATCCATCC TGCTCCATCT CTCTCCATAG AAGTCCATGG
GTGTCCATAC CGTTCCATCC GGAAGCATGT AACAGCATAG TACAGCATCA
GCCAGCATTA GTCAGCATGC CTGAGCATGG ACACGCATCG TCACGCATAA
25 TAGCGCATTG GCGCGCATCA ACTCGCATAT TAAGGCATCT GCAGGCATTT
GGCGGCATGA GAGGGCATAT AATGGCATCA TATGGCATAG CCTTGCATGG
GTAATCATTG ACGATCATGC CGGATCATTT ATGATCATTG CTAATCATCC
CGCCTCATCT ATCCTCATCA TTCCTCATAG TGTCTCATCC CCGGTCATGT
AGGGTCATTC GAATTCATCT TCATTCATTT TGCTTCATGA TAGTTCATAT
30 GATTTTCATAG ACTTTTCATTA ATTTTCATGC CGAAAGATCG TTAAAGATAT
TGCAAGATAC GTGAAGATCT CGTAAGATAT ATTAAGATAA TTAAAGATCG
GGACAGATTC GTCCAGATTT ATCGAGATCG TTCGAGATAA AGTGAGATAC
AAATAGATTA GCATAGATAG CACTAGATTC GAGTAGATTG ACGTAGATAA
TCGTAGATCG TATTAGATTT GCTTAGATCT GCAACGATGT GGCACGATTA
35 AGACCGATCG TGACCGATAA GGGCCGATCA AGTCCGATAT CTTCCGATAA

GGAGCGATTT GCCGCGATGA GTCGCGATTC CTGGCGATTG TTATCGATGG
AAGTCGATCC CGTTCGATGG GTCAGGATCG CTGAGGATCC GGTAGGATAC
ACACGGATGC CGACGGATTT ATACGGATTA CCCC GGATGA CTCCGGATTC
GTGCGGATTG TTTCGGATTT CGGGGGATCA ATGGGGATCT GTTGGGATAA
5 GACTGGATTT TCCTGGATCT CCTTGGATCC TAAATGATTA GCAATGATCA
ACGATGATCG TCGATGATAA AATATGATTT CTACTGATGG GACCTGATAA
ACCCTGATTG GGGCTGATGT TGTCTGATGG TACGTGATTT GCCGTGATCT
CATGTGATTC CGATTGATAA ATATTGATAT AGCTTGATAC ATTTTGATCG
TTTTTGATAA ACAAATATCC CCCAATATCA CAACATATAA AACCATATAC
10 GTCCATATGG CCGCATATTA CTGCATATGC CCAGATATCT CCTGATATAG
GGCTATATAA GAGTATATGT TATTATATGG AAAACTATAA TAAACTATCG
GCAACTATTG CACACTATAC GAGACTATAG ACGACTATTA ATGACTATGC
TATACTATAT CCACCTATCC ACCCCTATCA TCCCCTATAG GGAGCTATGG
AACGCTATAT AGGGCTATGA CTGGCTATGT TTATCTATTT TGCTCTATTC
15 ACTTCTATGC CGTTCTATTT ATTTCTATTA CATAGTATTA CGACGTATGG
GTGCGTATGT TTTCGTATGG TAAGGTATTC GCAGGTATCC TCGGGTATAC
AGATGTATAT CTATGTATAA GACTGTATGG ATCTGTATAC CAGTGTATGC
AGTTGTATCG TGTTGTATAA GGCATTATCG CGGATTATCC AATATTATGG
GTTATTATAC AGACTTATTA CTACTTATTT ACCCTTATGT CGCCTTATTC
20 GGGCTTATTG TGTCTTATTT TACGTTATGG AGGGTTATCT CTGGTTATCA
GGTGTATATAA GAATTTATTC TCATTTATCC CCCTTTATAG GCGTTTATAC
AGAAA ACTGA CTAAA ACTGT ACCAAA CTTT CGCAA ACTGC GGGAAA CTGG
TGTA AACTGT ACGCA ACTCT AATCA ACTTG GCTCA ACTAA GAAGA ACTAG
ACAGA ACTTA ATAGA ACTGC CCCGA ACTTC AGCGA ACTGT CTCGA ACTGA
25 AAGGA ACTAA TAGGA ACTCG GCGGA ACTTG GATGA ACTCT TCTGA ACTTT
CCGAC ACTCT CATA CACTTG CGACC ACTGA ATACC ACTGT CCCCC ACTTT
AGCCC ACTGC GATCC ACTCC TCTCC ACTTC TTTCC ACTGA ACGGC ACTCC
TGATC ACTCT CTATC ACTAG GACTC ACTGA TTCTC ACTCC GGGTC ACTAT
AGTTC ACTCA TGTT CACTAG CTTTC ACTCT AAAA GACTAG TAAAG ACTCA
30 GCAAG ACTTA GTAAG ACTGC GGCAG ACTGT GAGAG ACTAA ACGAG ACTTG
AATAG ACTCT ACCCG ACTCG TCCCG ACTAA GGAGG ACTGA TACGG ACTCT
GCCGG ACTTT AGGGG ACTGG CATGG ACTCC ATTTG ACTTG GTCATA CTTG
CCGATA CTGA CTGATA CTTC CGACT ACTCT ATACT ACTCA TTACT ACTAG
CTCCT ACTCC GTGCT ACTCG CGTCT ACTAG TTTCT ACTCT CGGGT ACTTA
35 ATGGT ACTTT TGATT ACTGA GACTT ACTCT TCCTT ACTTT CCTTT ACTTC

AGTTTACTGT CTTTACTGA CAGAACCTCT CCTAACCTTG CACCACCTTT
TATCACCTTC GCTCACCTCC TTAGACCTGG AAGGACCTCC CGTGACCTGG
GGATACCTTT GCCTACCTGA GTCTACCTTC CTGTACCTTG AACACCCTTT
TGCCCCCTGG AAGCCCCTCT GATCCCCTAA ACTCCCCTTG AAAGCCCTTA
5 GCAGCCCTAG CACGCCCTTC GAGGCCCTTG ACGGCCCTAA TCGGCCCTCG
TATGCCCTTT GCTGCCCTCT ATCTCCCTCG TTCTCCCTAA AGTTCCCTAC
TAAAGCCTAC CACAGCCTCG GAGAGCCTCC TCGAGCCTTC TTGAGCCTGA
CCACGCCTAG GACCGCCTTC TCCCGCCTCC CAGCGCCTTG CCTCGCCTCT
TGGGGCCTGC CATGGCCTAA ACATGCCTGT CGATGCCTTC AGCTGCCTTA
10 CTCTGCCTTT TCTTGCCTGA TTTTGCCTTC GGTATCCTTA GAACTCCTGG
ATACTCCTAC AGCCTCCTAT CTCCTCCTAA AAGCTCCTGA GCTGTCCTGA
GTTGTCCTTC CGCTTCCTGG GGGTTCCTGC AGAAAGCTAG TGAAAGCTCA
GACAAGCTGT GGGAAAGCTAA AGTAAGCTCT CTTAAGCTCA TCGCAGCTGT
GCTCAGCTGG GAAGAGCTGA TTAGAGCTCC TCGAGCTCT CTCGAGCTAG
15 AAGGAGCTGG GTGGAGCTAC CGTGAGCTCC CCGTAGCTCA TGGACGCTTC
CGACCGCTAG TTACCGCTCT TGCCCGCTCC GTGCCGCTAT CGTCCGCTCT
ATTCCGCTCA TTTCCGCTAG TCGGCGCTGC TTGGCGCTTA CCATCGCTTC
AGATCGCTGT CTATCGCTGA GACTCGCTAG ACCTCGCTTA ATCTCGCTGC
CAGTCGCTAC TGTTGCTGA AAAAGGCTGA CACAGGCTGC GAGAGGCTGG
20 ATGAGGCTAC TATAGGCTGT TCCCGGCTGG AGTCGGCTTG GGAGGGCTAG
AACGGGCTGT AGGGGGCTAA TGGGGGCTCG CTGGGGCTAT GGTGGGCTCT
ACTTGGCTAC CAAATGCTTG GCCATGCTAC CCGATGCTAG GAACTGCTCC
TCACTGCTTC TTAGTGCTGA TAGTGCTAC CGTCTGCTGA ATTCTGCTGT
AGATTGCTCA TGATTGCTAG CTATTGCTCT CGCTTGCTCC GGGTTGCTCG
25 TGTTTGCTCT CTTTGTAG CCAAATCTAC GACAATCTTG ACCAATCTAA
TCCAATCTCG CAGAATCTTC AAACATCTCA TAACATCTAG CACCATCTCC
GAGCATCTCG TCGCATCTTG TATCATCTCT GTCATCTTT ACAGATCTAT
AAGGATCTTT ACTGATCTCG TCTGATCTAA GGATATCTCC GTCTATCTCT
TGGTATCTAC CATTATCTGA CAAACTCTCA AACACTCTCC GGTACTCTGC
30 ACACCTCTAC CCCCTCTAA AAGCCTCTTC CACGCTCTCT TATGCTCTCC
GCTGCTCTTC GTTGCTCTGA AGATCTCTTG GGGTCTCTTA GCAAGTCTAT
GAGAGTCTTT TCGAGTCTCT TATAGTCTTG GCTAGTCTCG TGACGTCTGA
GACCGTCTCT TCCCGTCTTT CCTCGTCTTC AGTCGTCTGT CTTCGTCTGA
AACGGTCTTG GCCGGTCTAA CGATGTCTCT ATATGTCTCA TTATGTCTAG
35 CTCTGTCTCC GTGTGTCTCG CGTTGTCTAG TTTGTCTCT CAAATTCTGT

AGGATTCTAC ATTCTTCTTG AAAGTTCTGC GTAGTTCTAG CACGTTCTGA
CGGGTTCTAT ATGGTTCTAA TTGGTTCTCG GTTGTCTCT ACCTTTCTCG
TCCTTTCTAA GTAAAAGTCT GGCAAAGTCC CGGAAAGTCG TTGAAAGTAT
AATAAAGTGC GTTAAAGTAG CGCCAAGTTG GGGCAAGTTC TACGAAGTGC
5 GTCGAAGTCA CATGAAGTGT GAATAAGTTG ACATAAGTAA TCATAAGTCG
CCCTAAGTAC AAGTAAGTTA GCGTAAGTAG TCTTAAGTAT AGCACAGTTG
GCGACAGTGT TCTACAGTGG GGACCAGTCA TACCCAGTGT AGGCCAGTCG
TGGCCAGTAA CATCCAGTGC CTAGCAGTTG ACCGCAGTGG GGGGCAGTTT
TGTGCAGTTG AAATCAGTCC CACTCAGTCA TTGTCAGTGG GTTTCAGTGT
10 CTAAGAGTCC CGCAGAGTCT ATCAGAGTCA TTCAGAGTAG TGTAAGTCC
GTACGAGTTG ACGCGAGTGC CGGCGAGTTT ATGCGAGTTA CGAGGAGTCA
ATAGGAGTCT AGCGGAGTCC GATGGAGTGC ATTGGAGTAG TTTGGAGTCA
GCCTGAGTAT ACAATAGTCG TCAATAGTAA TAGATAGTTA GCGATAGTCA
ACTATAGTAT GGACTAGTGT AACCTAGTAG TACCTAGTCA GCCCTAGTTA
15 GTCCTAGTGC CTGCTAGTGG TCCGTAGTAT CCTGTAGTAC CACTTAGTGT
CGGTTAGTAA ATGTTAGTAT TATTTAGTGC GTTTTAGTCA GGCAACGTAA
GAGAACGTGT TATAACGTGG TGACACGTTA TCCCACGTGT CCTCACGTGC
AGTCACGTTT CTTCACGTTA GGAGACGTAT AACGACGTGG GTCGACGTAC
TGGGACGTCT CTGGACGTAG GGTGACGTGC ACATACGTCC CCCTACGTCA
20 TTAACCGTTT TGCACCGTTC ACTACCGTGC CGTACCGTTT ATTACCGTTA
GGACCCGTAC GTCCCGTAT TGGCCCGTCC TCCGCCGTGC TTCGCCGTTA
CCTGCCGTGT AGTGCCGTTT AAATCCGTAA TAATCCGTGC GCATCCGTTG
CACTCCGTAC GAGTCCGTAG ACGTCCGTTA ATGTCCGTGC TATTCCGTAT
AGAAGCGTAT CTAAGCGTAA GACAGCGTGG ATCAGCGTAC CAGAGCGTGC
25 AGTAGCGTCG TGTAAGGTAA TCGCGCGTGG GCTCGCGTGT CGAGGCGTAC
AGCGGCGTAA TGCGGCGTCG CTCGGCGTAT AAGGGCGTGT TTTGGCGTAC
CATTGCGTTA GAAATCGTTC TCAATCGTCC CCCATCGTAG GCGATCGTAC
CAACTCGTCG TACCTCGTAC CATCTCGTAT CCTGTCGTCA GGCTTCGTGC
CGGTTTCGTCC AATTTTCGTGG GTTTTCGTAC CACAAGGTCT TATAAGGTCC
30 GCTAAGGTTC GTTAAGGTGA CCACAGGTAT CAGCAGGTTT CCTCAGGTGC
CAAGAGGTCA AACGAGGTCC GGTGAGGTGC ACATAGGTGG CTCTAGGTTC
GCGTAGGTGA GTGTAGGTTC ACAACGGTAT AAGACGGTTT ACTACGGTGC
TCTACGGTAA AACCCGGTCT TGGCCGGTGG GGTCCGGTGT CCAGCGGTAC
GACGCGGTTC ACCGCGGTAA TCCGCGGTGC CAGGCGGTTC TAATCGGTGC
35 GTATCGGTCA ATGTCGGTGC TTGTCGGTAA GACAGGGTCC TCCAGGGTTC

TTCAGGGTGA CAGAGGGTCG CCTAGGGTTT AGTAGGGTGC GCACGGGTAC
CACCGGGTTG GAGCGGGTTC TCGCGGGTCC TTAGGGGTGT TGCGGGGTGC
ACTGGGGTTC CGTGGGGTGT ATTGGGGTGA GGATGGGTTG AGGTGGGTGA
CTGTGGGTTT TCAATGGTGG CAACTGGTGC AACCTGGTGA AGGCTGGTAT
5 CTGCTGGTAA GGTCTGGTCA TGAGTGGTTG CTTGTGGTTG GTATTGGTGT
GGCTTGGTGC CGGTTGGTGG AATTTGGTCC TCGAATGTGC TTGAATGTGA
GACCATGTGC ATCCATGTAG TTCCATGTCA CAGCATGTGG AGTCATGTCC
TGGGATGTTT ACATATGTTT CGATATGTGC AGCTATGTGA CTCTATGTGT
AAGTATGTAT GATTATGTCA TCTTATGTGA TTTTATGTGC GAAACTGTGA
10 TTAAGTGTCC TGCAGTGTCT CTCAGTGTAG AAGACTGTGG GTGACTGTAC
CGTACTGTCC TGGCCTGTTT GGTCTGTG AGAGCTGTAG TGAGCTGTCA
GACGCTGTGT GGGGCTGTAA AGTGCTGTCT CTTGCTGTCA GCTTCTGTAC
CCTAGTGTGG CACCGTGTGT CGGCGTGTAA ATGCGTGTAT TATCGTGTGC
GTTCGTGTCA TTAGGTGTTG CGTGGTGTG GGATGTGTGT AACTGTGTAG
15 TACTGTGTCA GCCTGTGTGA GTCTGTGTGC CTGTGTGTGG GAAATTGTCT
TCAATTGTTT TGCATTGTGA TAGATTGTAT GATATTGTAG ACTATTGTGA
ATTATTGTGC GCCCTGTAT TGAGTTGTGT GACGTTGTCA TCCGTTGTGA
TTCGTTGTGC AGTGTTGTGA CTTGTTGTGT GTATTTGTTG ACGTTTGTGC
CGGTTTGTAT ATGTTTGTGA ACTAAATTCC TGGCAATTGC CATCAATTAA
20 CCAGAATTAG GACGAATTTC TCCGAATTCC CAGGAATTG CCTGAATTCT
TAATAATTGG ATGTAATTCC TAAACATTAT GAGACATTCT TCGACATTTT
AATACATTAA TATACATTCC GCTACATTG GACCCATTTT TCCCCATTCT
CCTCCATTCC AACGCATTCC TACGCATTAA TGGGCATTGT GGTGCATTGG
ACATCATTGC CGATCATTTT ATATCATTTA CCCTCATTGA CTCTCATTTC
25 GTGTCATTTG TTTTCATTTT CAAAGATTGA AACAGATTTC CCGAGATTAA
TTACGATTGG AAGCGATTCC CGTCGATTGG CACGGATTTT TATGGATTTC
GCTGGATTCC AGATGATTCC TGATGATTAA GGGTGATTCA AGTTGATTAT
CTTTGATTAA AAAATATTGG GTAATATTAC GGCATATTAT GAGATATTGA
TTGATATTCC TGAATATTTT TCCCTATTGA TTCCTATTTT AGTCTATTGA
30 CTTCTATTTT GGAGTATTAA AGGGTATTAG TGGGTATTCA CCCTTATTCT
GATTTATTTT TCTTTATTCC TCAAACCTAT TAGAACTTTT GCGAACTTCT
GATAACTTTG ACTAACTTAA TCTAACTTCG GGACACTTGA TACCACTTCT
GCCCACTTTT AGGCACTTGG CATCACTTCC ACCGACTTCG TCCGACTTAA
AAATACTTGC GTATACTTAG CACTACTTGA CGGTACTTAT ATGTACTTAA
35 TTGTACTTCG GTTACTTCT GTAACCTTGT GGCACCTTGC CGGACCTTGG

AATACCTTCC CCTCCCTTAA CAAGCCTTAG TACGCCTTCC GCCGCCTTTC
GTCGCCTTGA CCGGCCTTTG CATGCCTTCT TCATCCTTGG TACAGCTTTG
GCCAGCTTCG CCGAGCTTCC GAACGCTTAG ACACGCTTTA ATACGCTTGC
CCCCGCTTTC AGCCGCTTGT CTCCGCTTGA AAGCGCTTAA TAGCGCTTCG
5 GCGCGCTTTG GATCGCTTCT TCTCGCTTTT ACGGGCTTCT AATGGCTTTG
GCTGGCTTAA TGATGCTTCC CGCTGCTTAG TTCTGCTTCT GGGTGCTTAC
CTTTGCTTCC TAAATCTTGC GTAATCTTCA ATGATCTTCG TTGATCTTAA
CCACTCTTGG ATCCTCTTTG GGAGTCTTCC GTCGTCTTCT TGGGTCTTAC
CATGTCTTGA ACATTCTTAT AAGTTCTTTT ACTTTCTTCG TCTTTCTTAA
10 CGAAAGTTTA ATAAAGTTTT CCCAAGTTGT AGCAAGTTTC TCTAAGTTGC
TTTAAGTTTA AGGCAGTTCC CATCAGTTGG CCAGAGTTGA CTAGAGTTTC
ACCGAGTTGC CGCGAGTTTT ATCGAGTTTA TGTGAGTTTC AAATAGTTTCG
TAATAGTTAA GAGTAGTTCA TCGTAGTTTA TTGTAGTTGC AATTAGTTAT
GGCACGTTTCG CGGACGTTCC AATACGTTGG GTTACGTTAC AGACCGTTTA
15 CTACCGTTTT ACCCCGTTGT CGCCCGTTTC GGGCCGTTTG TGTCCGTTTT
TACGCGTTGG AGGGCGTTCT CTGGCGTTCA GGTGCGTTAA GAATCGTTTC
TCATCGTTCC CCCTCGTTAG GCGTCGTTAC GCCAGGTTGC GTCAGGTTTA
CCGAGGTTGG ATACGTTTCG TTACGTTTAA TAGCGGTTGC GTGCGGTTCA
CGTCGGTTAA ATTCGGTTAT GTAGGGTTTT GGCGGGTTTC CGGGGGTTTTG
20 TGATGGTTGG CTTTGGTTGG AAAATGTTAA TAAATGTTTCG GCAATGTTTG
CACATGTTAC GAGATGTTAG ACGATGTTTA ATGATGTTGC TATATGTTAT
CCACTGTTCC ACCCTGTTCA TCCCTGTTAG GGAGTGTTGG AACGTGTTAT
AGGGTGTTGA CTGGTGTTGT TTATTGTTTT TGCTTGTTTC ACTTTGTTGC
CGTTTGTTTT ATTTTGTTTA GAAAATTTAC ATAAATTTGG CCCAATTTTG
25 TAGAATTTCC GCGAATTTTC GTGAATTTGA TACCATTTTC GCCCATTTCC
CCGCATTTTCG CATCATTTTT CGCGATTTGG GGGGATTTGC GCTTATTTGA
GTTTATTTTC TAAACTTTTA GCAACTTTCA ACGACTTTTCG TCGACTTTAA
AATACTTTTT CTACCTTTGG GACCCTTTAA ACCCCTTTTG GGGCCTTTGT
TGTCCTTTGG TACGCTTTTT GCCGCTTTCT CATGCTTTTC CGATCTTTAA
30 ATATCTTTAT AGCTCTTTAC ATTTCTTTTCG TTTTCTTTAA CAAAGTTTAT
CCGAGTTTTT AGGAGTTTGC CATAGTTTCG CCCCCTTTCT GATCGTTTTC
TCTCGTTTCC AAAGGTTTAC GTAGGTTTGG CACGGTTTAA ACGGGTTTTC
CGGGGTTTTGT ATGGGTTTGA TGATGTTTTT TCCTGTTTGA TTCTGTTTTC
AGTTGTTTTA CTTTGTTTTT GCAATTTTGT GGCATTTTAA AGACTTTTCG
35 TGACTTTTAA GGGCTTTTCA AGTCTTTTAT CTCTTTTAA GGAGTTTTTT

GCCGTTTTGA GTCGTTTTTC CTGGTTTTTG TTATTTTTGG AAGTTTTTCC
CGTTTTTTGG

[0073] The foregoing code set is also applicable to the following dataset for Reed-Solomon
5 + 2 base, 256 sequence, 1 base error correction, 2 nocall correction:

5 AAAAAAA CCAAACA GCAAATG GGAAAGA TTAAATA AACAAAGT CACAACC
ACCAAAC TCCAACG GGCAAAT CTCAATG GTCAACA GAGAAGG AGGAAAG
TGGAACC TATAATT AGTAACA ATTAAAT AAACACC CAACAGT TAACAAG
CCACAAC ATACATG CACCAAA ACCCACA GTCCAAC CCGCATT TGGCAAA
TTGCACT GATCAGC GCTCAAT AGTCAAC TGTCACG CTTCAGA AAAGAGG
10 TCAGACT GGAGAAG CTAGATT GCCGATT AGCGATC GAGGAAA ACGGATG
AGGGAGA CTGGAAC CATGATA GATGACG TGTGAGC GTTGATC AAATATT
TGATACA CTATAGG TTATAAT GCCTAGG CGCTACG GGCTATA TTCTAGA
AAGTAAC TAGTACG ACGTAGT CCGTACC GGGTAGC TTGTATC TATTA
ATTTATA CAAACAC ACAACCC CCAACGT TCAACAG GGAACCT GTAACAA
15 AACACCA CCCACAA TGCACCT GTCACGT CAGACTT ATGACGG GATACAT
GCTACGC AGTACGT CGTACCC CAACCCA GAACCTG ACACCAA GTACCCC
AACCCAC TACCCCG ACCCCGT CCCCCC GGCCCGC TTCCCTC GCGCCGG
CGGCCCC GGGCCTA TTGCCGA TCTCCTT CGTCCAA ATTCCGC CTTCCT
TCAGCGA TGAGCTG GTAGCGG TACGCGC CCCGCGG TCCGCAT CGCGCTA
20 GGCGCCG TAGGCTA GCGGCC AGGGCCT CGGGCGC ATGGCAA AATGCTC
ATTGCCG TTTGCAC CGATCAG TGATCGT GTATCTT CCCTCTT TGCTCAA
TTCTCCT CAGTCAA ACGTCCA GTGTAC CATTCCG GATTCTA ACTTCAG
TCTTCCC CTTTCTC GAAAGAG ACAAGTA AGAAGGG AACAGTC ATCAGCG
TTCAGAC AAGAGGA GGGAGAA GTGAGCT TATAGGC CCTAGGG TCTAGAT
25 CGTAGTA GGTAGCG CAACGTC TCACGTG TGACGGA CTACGCG GTACGTA
CCCCGTA GCCCGCG CGCCGGG TGCCGAT TTCCGCA CAGCGAT TAGCGGG
ACGCGCT CCGCGGC GGGCGCC GATCGTT GAAGGGA AGAGGAA ATAGGCT
CTAGGGC GACGGAT GCCGGGC AGCGGGT CGCGGCC AAGGGAG TAGGGCC
CCGGGCG GCGGGTA GGGGGGG TTGGGTG AATGGCA CCTGGAA TGTGGTT
30 GTTGGGT AAATGGC CAATGCT ACATGAT GGATGAC ACCTGGA AGCTGTG
ACGTGTC GGGTGTT CTGTGAG TTGTGGT GATTGCC TCTTGTA CGTTGAT
TGTTGGG CTTTGCA GTTTGTG CAAATGG TAAATAT TCAATGC ATAATTT
TACATGA CGCATTC TAGATTC CCGATTG GCGATCA CGGATGA ATGATAC
TTGATCG AATATTA GCTATAG TGTATCT TTTATAA GCACTTT AGACTTC

AACCTGG TCCCTCT GGCCTAG CTCCTTT GAGCTGT GCGCTAC AGGCTAT
 ATGCTCA TATCTTG CCTCTTC ATTCTAG TTTCTCC CAAGTAA ACAGTCA
 GTAGTAC AACGTCC CACGTGT TACGTAG CCCGTAC ATCGTTG CGGGTAG
 TGGGTGT GTGGTTT GCTGTGA CGTGTCA GGTGTTG CTTGTAT TTTGTGG
 5 TAATTTA GCATTCC AGATTCT CGATTGC ATATTAA CACTTTG GACTTCA
 TCCTTTC ATCTTGT CTCTTCC TCGTTGA TGGTTTG GTGTTGG AATTTAT
 ACTTTGC CCTTTCT GGTTTGT TTTTTTT

10 **[0074]** The error correction technique of the invention can be used in many applications.
 The technique will correct a base call 44 to a corrected base call 46, where the bases could
 part of a model system or a construct made of a tag and another sequence which could be an
 unknown sequence from a defined set. Once the DNA markers, for example DNA tags, are
 fully recovered by these methods, the population of identified bases can then be assembled to
 15 recover and correct sequence information for the target nucleic acid and/or identify the
 presence of particular sequences in the target nucleic acid. There is also potential for
 exploiting the higher accuracy of tag recovery in developing more efficient and potentially
 faster genome sequencing techniques, such as long fragment read (LFR) techniques. In some
 embodiments, the identified bases are assembled into a complete sequence through alignment
 20 of overlapping sequences obtained from multiple sequencing cycles performed on multiple
 DNBs. As used herein, the term “complete sequence” refers to the sequence of partial or
 whole genomes as well as partial or whole target nucleic acids.

[0075] Long Fragment Reads (LFR) technology enables independent sequencing and
 analysis of the two parental chromosomes in a diploid sample. See for example, US
 25 Published Patent Application US 2007/0072208 corresponding to PCT Published Patent
 Application WO2006/138284 published 28 December 2006. As applied to the sequencing of
 polyploidy organisms, such as diploid human genomes, LFR allows heterozygote phasing
 over large intervals (potentially entire chromosomes), even in areas with high recombination
 rates. In addition, by distinguishing calls from the two chromosomes, LFR allows higher
 30 confidence calling of homozygous positions (>99% of the genome) at low coverage.
 Additional applications of LFR include, but are not limited to, resolution of extensive
 rearrangements in cancer genomes and full-length sequencing of alternatively spliced
 transcripts.

[0076] According to a typical application of LFR to human genomic DNA, genomic DNA of approximately 100 kbp is used as the input for LFR, as the length of input DNA impacts the interval over which phasing can be performed. This high molecular weight genomic DNA is aliquotted into a 384 well plate such that approximately 0.1 haploid genomes (10%
5 of a haploid genome) are aliquotted into each well. The DNA fragments in each well are amplified, and this amplified DNA is fragmented to about 500bp. The DNA in each well is ligated to adaptor arms containing a unique identifier, and the ligated DNA from all 384 wells is then pooled into a single tube.

[0077] This pooled DNA is then used as input to a standard library construction (such as
10 that developed by Complete Genomics, Inc. of Mountain View, California) and sequencing processes. In aggregate, the 384 wells will contain approximately 40 fragments, spanning each position in the genome, with about 20 fragments coming from the maternal chromosome and 20 from the paternal chromosome. At a rate of 0.1 genome equivalents per well, there is a 10% chance that fragments in a well will overlap, and a 50% chance that any such
15 overlapping fragments are derived from separate parental chromosomes. Thus, approximately 95% of the data from a well will be derived from a single parental chromosome.

[0078] In order to resolve the parental chromosomes, the reads from each well are effectively assembled independently from other wells. The data is then mapped to one or
20 more reference genomes, and the reads that map near each other are grouped by their unique identifiers, enabling reconstruction of the approximate 100kbp haploid fragments in each well. Single nucleotide polymorphisms (SNPs) within the sample are then used to distinguish between 100kbp fragments from the maternal and paternal chromosomes.

[0079] The initial 40 genome equivalents described above yield on average a 100kbp
25 maternal fragment starting every 5kbp and a 100kbp paternal fragment every 5kbp. Thus, two consecutive maternal fragments will overlap each other on average by about 95 kbp. In the human genome, there are typically 50-150 Single Nucleotide Polymorphisms (SNPs) within 95 kbp, many of which will be heterozygous in any given sample.

[0080] Using these SNPs, maternal fragments are distinguished from paternal fragments.
30 By chaining together overlapping fragments; large maternal and paternal segments (up to complete chromosomes) can be constructed separately. Phasing will not be possible across long repeat sections such as satellites in centromeric regions. But for most practical purposes LFR increases effective read length from 35 bp to over 100 kbp.

[0081] In further embodiments, assembly methods utilize algorithms that can be used to “piece together” overlapping sequences to provide a complete sequence. In still further embodiments, reference tables are used to assist in assembling the identified sequences into a complete sequence. A reference table may be compiled using existing sequencing data on the organism of choice. For example human genome data can be accessed through the National Center for Biotechnology Information at ftp.ncbi.nih.gov/refseq/release, or through the J. Craig Venter Institute at <http://www.jcvi.org/research/huref/>. All or a subset of human genome information can be used to create a reference table for particular sequencing queries. In addition, specific reference tables can be constructed from empirical data derived from specific populations, including genetic sequence from humans with specific ethnicities, geographic heritage, religious or culturally-defined populations, as the variation within the human genome may slant the reference data depending upon the origin of the information contained therein.

[0082] In any of the embodiments of the invention discussed herein, a population of nucleic acid templates and/or DNBs may comprise a number of target nucleic acids to substantially cover a whole genome or a whole target polynucleotide. As used herein, “substantially covers” means that the amount of nucleotides (i.e., target sequences) analyzed contains an equivalent of at least two copies of the target polynucleotide, or in another aspect, at least ten copies, or in another aspect, at least twenty copies, or in another aspect, at least 100 copies. Target polynucleotides may include DNA fragments, including genomic DNA fragments and cDNA fragments, and RNA fragments. Guidance for the step of reconstructing target polynucleotide sequences can be found in the following references: Lander et al, *Genomics*, 2: 231-239 (1988); Vingron et al, *J. Mol. Biol.*, 235: 1-12 (1994); and like references.

[0083] Reed-Solomon (RS) code-based sensors in the form of DNB, beads and the like, may be spiked into into the mix of conventional genomic sensors on a substrate. For instance, one could construct a substrate sensor set with a mixture of 99.9% conventional genomic sensors and 0.1% RS sensors. After reading a series of bases, e.g., 10, one could select from the mix candidate RS sensors with high fidelity. The candidate RS sensors will be those that differ from the RS codes by at least K bases, where K is normally 0 or 1. These candidate RS sensors can be delivered to the RS decoding algorithm for the characterization of the system. A 0.1% spiking of RS in a lane of genomic information will provide sufficient information to enable RS analysis while providing minimal contamination to the genomic sequences, as a library would contain multiple RS codes in a DNA short-read. Other spiking densities are also contemplated, as for example as much as ten per cent but preferably in the

range of one per cent to five per cent. The set that is spiked in is preferably a subset of the 4096 possible RS codes that are available. This subset is preferably optimized to have minimal hits to the genomic DNA. By doing so, no extra space needs to be set aside on the substrate for the RS sensor system. Moreover, for each lane of the genomic experiment, an internal control would be available, i.e., the RS sensors that are spiked in. Alternatively, the spiked-in RS sensors can have an independent short tag that would differentiate between them and the genomic data.

[0084] Figure 2 illustrates results of the use of Reed-Solomon correction on a set of 7-mer DNA sequences. Column 1 is an index. Column 2 is the calls of the observed sequence. Column 3 is the corrected sequence after error correction according to the invention. Column 4 indicates whether a call was recovered, i.e., changed from the observed call. Columns 5 through 11 are the scores at the seven positions for each of the bases in the seven-base sequence, i.e., the estimated likelihood that the original call was correct. The error in the scores could be a function of problems in either physical or optical interference between adjacent interrogation sites (cross-talk), a weak DNB at an interrogation site, a missing DNB at an interrogation site allowing a reading to be taken based on an adjacent interrogation site, problems in stripping, washing and flushing between sequencing cycles and errors in sequencing. In the course of processing and correction of five sequences, two bases were recovered. As a condition, the processing algorithm was blind to all scores and to all other sequences (when processing each sequence).

[0085] One of the outcomes of Reed-Solomon decoding of the genomic data is that certain sites do not result in a correction of a call. The no-call output is typically an indication that the interrogation site has more than one error, which is often the case when the interrogation site is empty and/or the report of an observed sequence is based on readings from adjacent interrogation sites (cross-talk). Thus the Reed-Solomon decoding actually identifies bad interrogation sites. Figure 3 is a table summarizing an experiment comparing observed sequences and Reed-Solomon corrected sequences. Column 1 is the categories of A. Valid Sequences, B. Uncalled Sequences, C. Invalid Sequences, and D. Total Sequences accounted for. There was a total of 50625 7-mer sequences accounted for. Of that, 20647 were observed valid sequences and 29978 invalid sequences. After Reed-Solomon correction, the number of valid sequences increased to 35317 and 15308 sequences were designated as no-calls, eliminating all invalid sequences from the set. The result is that the Reed-Solomon correction technique was able to recover about 70% of the valid DNBs and suppress the empty interrogation sites and low-quality DNBs by designating them as no-calls.

[0086] The inventive method can operate on blocks of delimited sets of monomer subunits that are artificial DNA sequences. The artificial DNA sequences are made up of oligonucleotides, or at least one oligonucleotide. The method according to the invention works on real systems containing tags, as well as in model systems where the model content is known and used to characterize real systems. Tags may be employed in Long Fragment Read techniques. Tags are useful particularly in multiplexing a plurality of individual samples, in multiplexing a plurality of tissues of a sample, and in multiplexing multiple libraries of a sample.

[0087] In the preparation of samples wherein the monomer subunit sequence has tags, the physical elements containing the tags can be randomly placed on a substrate, the tags being used as markers to identify both location and type of subunit sequence.

[0088] It should be understood that the delimited set of monomer subunits referred to herein encompasses a sequence of nucleotides and/or a sequence of amino acids, both natural and artificial, and oligonucleotides, both natural and artificial.

[0089] The methodology thus improves the accuracy and efficiency of DNA sequencing for a model system by increasing the true positives and true negatives, and decreasing the number of false positives and false negatives. It is also useful as a check against other sensor systems. There are numerous applications for fully-recoverable set of DNA tags. In addition to improving the accuracy of genome mapping (not applicable to genome mapping), if it were known that the DNA tags were accurate, then the other sources of error or imperfections in the analysis and preparation processes could be better identified and the errors and improvements could be quantified. Furthermore, other methods of error correction verification can be verified and calibrated. Still further, various techniques for verification and error correction can be combined to provide even higher efficiency and accuracy in the sequencing process.

[0090] Figure 4 is a representative figure similar to Figure 2 intended to depict 10-base codes that are a preferred embodiment.

[0091] Figure 5 depicts a generalized apparatus 100 incorporating and operative according to another aspect of the invention. The apparatus 100 is for processing errors in data used for identifying a sequence of an oriented linear heteropolymer, such as a polynucleotide, a oligonucleotide, an amino acid and a DNA sequence, including both natural and artificial sequence types. Processing includes either identifying one or more errors in the sequence or correcting one or more errors in the sequence. The apparatus includes a subsystem 101

operative to identify the monomer subunit sequence in order to obtain preliminary codes 44, the preliminary codes intended to resemble members of an expected known delimited code set of block codes. The sequence identifier may identify as little as one preliminary code or sequence but it is preferable to identify several sequences or codes. There may be an optional
5 analysis subsystem 102 for preliminary analysis of the preliminary codes 44 using a computer processing system, in order to obtain (analyzed) candidate codes 46. The candidate codes are stored in a table 106 in a storage subsystem 104 to expedite the processing of several sequences or codes together. The candidate codes or sequences 46 are retrieved from the storage subsystem 104 and provided to an error correction/detection, a processing subsystem
10 108 for detecting errors or invoking error correction. The output of subsystem 108 is an output code 114. The output code may take the form of the sequence or sequence listing that has been corrected according to the invention or a set of indicators of error content for each sequence, such as zero, one or two errors, or a no-call, indicating that the sequence is not identifiable. This system can be used for modeling, tagging, spiking of natural sequences for
15 identification, and for error correction to correct for errors in the initial (visual) readings and, for model systems, sequence calling for a known system..

[0092] According to the invention, compositions may be formed, such as for use in modeling and improved genome sequencing, as well as other applications. Compositions may comprise linear oriented heteropolymers of blocks of monomer subunit sequence having
20 coding of an expected known delimited code set of block codes, and specifically Reed-Solomon codes, namely codes produced according to the Reed-Solomon algorithm, and specifically those of a length between 5 and ten monomer subunits, wherei the ratio of distance between codes and length of code is at least 20 percent. Specific heteropolymer components may include polypeptides, polynucleotides, such as DNA nanoballs,
25 oligonucleotides Such a composition may include and be attached to a substrate wherein the heteropolymers are attached at spaced apart interrogation sites, and the sites of attachment may be ordered or randomly arranged.

[0093] By way of summary, the features of the invention include the following:

[0094] One embodiment of the invention is a method for processing errors in data
30 corresponding to a sequence of an oriented linear heteropolymer comprising monomers involving acquiring an output of data from an apparatus operative to identify a sequence or a subsequence within the linear heteropolymer to obtain preliminary codes, the preliminary codes intended to resemble members an expected known delimited code set of block codes; applying the preliminary codes as coefficients of a vector of symbols of the delimited ode set;

invoking, using a computer processing system, an error detection or correction algorithm for at least block error identification on the vector to at least identify errors in the symbols; and outputting at least error identification to a storage subsystem for utilization.

[0095] The method may further include analyzing the preliminary codes against the
5 expected known delimited code set, for each position of the preliminary codes associated with positions of the sequence, to obtain a candidate code set; and assembling the candidate code set into a table of candidate codes in a storage element for the applying step.

[0096] The basic method may further include correcting each correctable error in the preliminary codes to match with the delimited set to obtain a corrected sequence.

10 [0097] The foregoing improvement on the method may further include the step of generating a code set for encoding the symbols, wherein each random pair from the code set is guaranteed to have a minimum ratio R of distance D to length L of greater than 20 percent from every other member of the set.

[0098] In the foregoing improvement on the method, the distance D may be greater than or
15 equal to 3 for length L equal to 10.

[0099] Moreover, in the foregoing improvement on the method, the distance D may be greater than or equal to 3 for length L equal to 7.

[0100] In the basic method, the correction code set may be a subset of blocks D of N members of a Reed-Solomon code set, wherein N is greater than D .

20 [0101] In the basic method, the error detection or correction algorithm may be a Reed-Solomon algorithm.

[0102] In the method, the block of the delimited set may be an artificial DNA sequence comprising an oligonucleotide.

[0103] In the method, each block of the delimited set of monomer subunits may include a
25 tag, wherein the tag is used in a Long Fragment Read technique.

[0104] In the method, each block of the delimited set may include a tag, and there may be a further step of using the tag in multiplexing a plurality of individual samples.

[0105] In the method, each block of the delimited set may include a tag, and there may be a further step of using the tag in multiplexing a plurality of tissues of a sample.

[0106] In the method, each block of the delimited set may includes a tag, and there may be a further step of using the tag in multiplexing multiple libraries of a sample.

[0107] In the method wherein the block comprises tags, elements containing the tags may be randomly placed on either a patterned or an unpatterned substrate.

5 [0108] In the method, the block of the delimited set may be a sequence of nucleotides.

[0109] In the method, the block of the delimited set is a sequence of amino acids.

[0110] In the foregoing improvement of the method, the amino acids may be artificial.

[0111] The basic method may further include the step of generating a code set for encoding the symbols, wherein each random pair from the code set is guaranteed to have a minimum
10 ratio R of distance D to length L of greater than 20 percent from every other member of the set.

[0112] In the foregoing improvement on the method, the distance D may be greater than or equal to 3 for length L equal to 10.

[0113] In the foregoing improvement on the method, the distance D may be greater than or
15 equal to 3 for length L equal to 7.

[0114] In the basic method, the correction code set may be a subset of blocks D of N members of a Reed-Solomon code set, wherein N is greater than D.

[0115] In the basic method, there may be the further step of designating as no-calls those preliminary corrected code sequences having at least one error.

20 [0116] In a further embodiment, the invention is a method for processing errors in data corresponding to a sequence of bases of a nucleic acid construct involving the steps of acquiring a set of data representative of the nucleic acid construct of interest from molecular attachment sites; analyzing, for each code of a delimited set of codes using a computer processing system, from the series of data, labels associated with positions of the molecular
25 attachment sites to obtain preliminary base calls for the sequence of interest; assembling the preliminary base calls into an array of preliminary base call sequences; applying the preliminary base call sequences as coefficients of polynomials; invoking a Reed-Solomon algorithm, using a computer processing system, for detection of errors on the polynomial to at least identify errors in the polynomial; and outputting said at least said identified errors to a
30 storage device.

[0117] In the basic error processing method, the sequence of bases of the nucleic acid construct may be an artificial DNA sequence comprising an oligonucleotide.

[0118] In the basic error processing method, each sequence of bases of the nucleic acid construct may include a tag, wherein the tag is used in Long Fragment Read techniques.

5 [0119] In the basic error processing method, each sequence of bases of the nucleic acid construct may include a tag and further include the step of using the tag in multiplexing a plurality of individual samples.

[0120] In the basic error processing method, each sequence of bases of the nucleic acid construct may include a tag and further include the step of using the tag in multiplexing a
10 plurality of tissues of a sample.

[0121] In the basic error processing method, each sequence of bases of the nucleic acid construct may include a tag and further include the step of using the tag in multiplexing multiple libraries of a sample.

[0122] In the basic error processing method, the sequence may comprise tags and elements
15 containing the tags may be randomly placed on a substrate.

[0123] In the basic error processing method, the sequence of bases of the nucleic acid construct may be a sequence of nucleotides.

[0124] The basic error processing method may further include the step of generating a correction code set wherein each random pair from the code set is guaranteed to have a
20 minimum ratio R of distance D to length L of greater than 20 percent from every other member of the set.

[0125] In the foregoing improvement, the distance D may be greater than or equal to 3 for length L equal to 10.

[0126] In the foregoing improvement, the distance D may be greater than or equal to 3 for
25 length L equal to 7.

[0127] In the basic error processing method, the correction code set may be a subset of blocks D of N members of the Reed-Solomon code set, wherein N is greater than D .

[0128] The basic error processing method may further include designating as no-calls those preliminary corrected code sequences having at least one error.

[0129] In the basic error processing method, the outputting may include no-call sequences and no-calls within a sequence.

[0130] The basic error processing method may further include correcting each single error in the sequence of delimited set preliminary base calls in the preliminary base call sequence
5 to obtain corrected base call sequences.

[0131] In a further embodiment of the invention, an apparatus for processing errors in data corresponding to a monomer subunit sequence of an oriented linear heteropolymer including a nucleic acid involves a subsystem apparatus, coupled to an apparatus operative to identify the monomer subunit sequence, for obtaining preliminary codes corresponding to the
10 monomer subunit sequence, the preliminary codes being candidates of a delimited set of codes; a storage subsystem for assembling the candidates into a table of candidate corrected codes; an error processing subsystem for invoking, using a computer processing system, error correction or detection on the candidates to at least identify errors or for correcting each correctable error in a block of the delimited set; and an output subsystem for receiving output
15 of the error processing system.

[0132] In the basic apparatus, the error correcting code may be a Reed-Solomon code.

[0133] In the basic apparatus, the block of the delimited set may be a sequence of nucleotides.

[0134] In the basic apparatus, the block of the delimited set may be a sequence of amino
20 acids.

[0135] In the apparatus, the amino acids may be artificial.

[0136] In the basic apparatus, the block of the delimited set may be an artificial DNA sequence comprising an oligonucleotide.

[0137] In a further embodiment of the invention, an apparatus for correcting errors in data
25 identifying a sequence of bases of a nucleic acid construct involves an acquisition system for acquiring a series of representations for a set of the nucleic acid construct of interest from genome attachment sites; a first data processing unit for analyzing, from the series of representations, labels associated with positions of the genome attachment sites to obtain a preliminary base call of a delimited set of codes; for repeating the analyzing for each code in
30 the set of codes for the sequence of interest; and for assembling the preliminary base calls into a table of preliminary base call sequences; and a second data processing unit for applying the preliminary base call sequences as coefficients in the form of a Reed-Solomon code, for

invoking Reed-Solomon error correction to at least identify errors, and for correcting each single error in a block of preliminary base calls in the preliminary base call sequence to obtain corrected base call sequences; and a data utilization system for receiving the corrected base call sequences.

- 5 [0138] In the error correcting apparatus, the second data processing subsystem may be a mechanism for comparing the preliminary base call sequences against an expected set of base call sequences.

[0139] In the error correcting apparatus, the acquisition subsystem may be an image acquisition system and the representations may be images.

- 10 [0140] In a further embodiment according to the invention, a composition may include a plurality of isolated oriented linear heteropolymers, each heteropolymer having a preselected block of a monomer subunit sequence, the monomer subunit sequence having a coding corresponding to at least one Reed-Solomon code of a length corresponding to five to ten monomer subunits, for use in processing errors in sequencing an isolated oriented linear
15 heteropolymer.

[0141] In the composition, each randomly selected pair of the heteropolymers of the set may have a minimum ratio R of distance D to length L of 20 percent.

[0142] In the composition, the heteropolymer may be a polypeptide.

[0143] In the composition, the heteropolymer may be a polynucleotide.

- 20 [0144] In the composition, less than five percent of the heteropolymers may comprise a preselected block of monomer subunit sequence.

[0145] The composition may further include a substrate for attachment of a plurality of the heteropolymers.

- [0146] In the foregoing version of the composition, the heteropolymers may be attached to
25 the substrate at spaced apart interrogation sites.

[0147] In the composition, heteropolymers may be DNA nanoballs.

- [0148] The present specification provides a complete description of the methodologies, systems and/or structures and uses thereof in example aspects of the presently-described technology. Although various aspects of this technology have been described above with a
30 certain degree of particularity, or with reference to one or more individual aspects, those

skilled in the art could make numerous alterations to the disclosed aspects without departing from the spirit or scope of the technology hereof. Since many aspects can be made without departing from the spirit and scope of the presently described technology, the appropriate scope resides in the claims hereinafter appended. Other aspects are therefore contemplated.

- 5 Furthermore, it should be understood that any operations may be performed in any order, unless explicitly claimed otherwise or a specific order is inherently necessitated by the claim language. It is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative only of particular aspects and are not limiting to the embodiments shown. Unless otherwise clear from the context or expressly
- 10 stated, any concentration values provided herein are generally given in terms of admixture values or percentages without regard to any conversion that occurs upon or following addition of the particular component of the mixture. To the extent published references and patent documents referred to in this disclosure are necessary for support of the claimed invention, they are to be considered a part of this disclosure. Changes in detail or structure
- 15 may be made without departing from the basic elements of the present technology as defined in the following claims.

WHAT IS CLAIMED IS:

- 1 1. A method for processing errors in data corresponding to a sequence of
2 an oriented linear heteropolymer comprising monomers, the method comprising:
3 acquiring an output of data from an apparatus operative to identify a sequence
4 or a subsequence within the linear heteropolymer to obtain preliminary codes, the preliminary
5 codes intended to resemble members an expected known delimited code set of block codes;
6 applying the preliminary codes as coefficients of a vector of symbols of the
7 delimited ode set;
8 invoking, using a computer processing system, an error detection or correction
9 algorithm for at least block error identification on the vector to at least identify errors in the
10 symbols; and
11 outputting at least error identification to a storage subsystem for utilization.
- 1 2. The method according to claim 1 wherein each block of the delimited
2 set of monomer subunits includes a tag, further including the step of using the tag in
3 multiplexing a plurality of tissues of a sample.
- 1 3. The method according to claim 1 wherein each block of the delimited
2 set of monomer subunits includes a tag, further including the step of using the tag in
3 multiplexing multiple libraries of a sample.
- 1 4. The method according to claim 1 wherein the monomer subunit
2 sequence comprises tags and wherein elements containing said tags are randomly placed on a
3 substrate.
- 1 5. The method according to claim 1 wherein the block of the delimited set
2 of monomer subunits is a sequence of nucleotides.
- 1 6. The method according to claim 1 wherein the block of the delimited set
2 of monomer subunits is a sequence of amino acids.
- 1 7. The method according to claim 6 wherein the amino acids are artificial.
- 1 8. The method according to claim 1 further including the step of
2 generating a correction code set wherein each random pair from the code set is guaranteed to
3 have a minimum ratio R of distance D to length L of greater than 20 percent from every other
4 member of the set.

1 9. The method according to claim 8 wherein the distance D greater than
2 or equal to 3 for length L equal to 10.

1 10. The method according to claim 8 wherein the distance D greater than
2 or equal to 3 for length L equal to 7.

1 11. The method according to claim 1 wherein said correction code set is a
2 subset of blocks D of N members of a Reed-Solomon code set, wherein N is greater than D.

1 12. The method according to claim 1 further including designating as no-
2 calls those preliminary corrected code sequences having at least one error.

1 13. A method for processing errors in data corresponding to a sequence of
2 bases of a nucleic acid construct comprising:
3 acquiring a set of data representative of the nucleic acid construct of interest
4 from molecular attachment sites;
5 analyzing, for each code of a delimited set of codes using a computer
6 processing system, from the series of data, labels associated with positions of the molecular
7 attachment sites to obtain preliminary base calls for the sequence of interest;
8 assembling the preliminary base calls into an array of preliminary base call
9 sequences;
10 applying the preliminary base call sequences as coefficients of polynomials
11 invoking a Reed-Solomon algorithm, using a computer processing system, for
12 detection of errors on the polynomial to at least identify errors in the polynomial; and
13 outputting said at least said identified errors to a storage device.

1 14. The method according to claim 13 further including the step of
2 generating a correction code set wherein each random pair from the code set is guaranteed to
3 have a minimum ratio R of distance D to length L of greater than 20 percent from every other
4 member of the set.

1 15. The method according to claim 14 wherein the distance D greater than
2 or equal to 3 for length L equal to 10.

1 16. The method according to claim 13 further including correcting each
2 single error in the sequence of delimited set preliminary base calls in the preliminary base
3 call sequence to obtain corrected base call sequences; and
4 outputting the corrected base call sequences to a storage device.

1 17. An apparatus for processing errors in data corresponding to a monomer
2 subunit sequence of an oriented linear heteropolymer including a nucleic acid comprising:
3 a subsystem apparatus, coupled to an apparatus operative to identify the
4 monomer subunit sequence, for obtaining preliminary codes corresponding to the monomer
5 subunit sequence, the preliminary codes being candidates of a delimited set of codes;
6 a storage subsystem for assembling the candidates into a table of candidate
7 corrected codes;
8 an error processing subsystem for invoking, using a computer processing
9 system, error correction or detection on the candidates to at least identify errors or for
10 correcting each correctable error in a block of the delimited set; and
11 an output subsystem for receiving output of said error processing system..

1 18. The apparatus according to claim 17 wherein the block of the delimited
2 set of monomer subunits is an artificial DNA sequence comprising an oligonucleotide.

1 19. A composition comprising:
2 a plurality of isolated oriented linear heteropolymers, each heteropolymer
3 comprising a preselected block of a monomer subunit sequence, the monomer subunit
4 sequence having a coding corresponding to at least one Reed-Solomon code having a length
5 corresponding to five to ten monomer subunits, wherein each randomly selected pair of said
6 heteropolymers of the set has a minimum ratio R of distance D to length L of 20 percent.

1 20. The composition according to claim 19 wherein the heteropolymers are
2 polynucleotides in the form of DNA nanoballs.

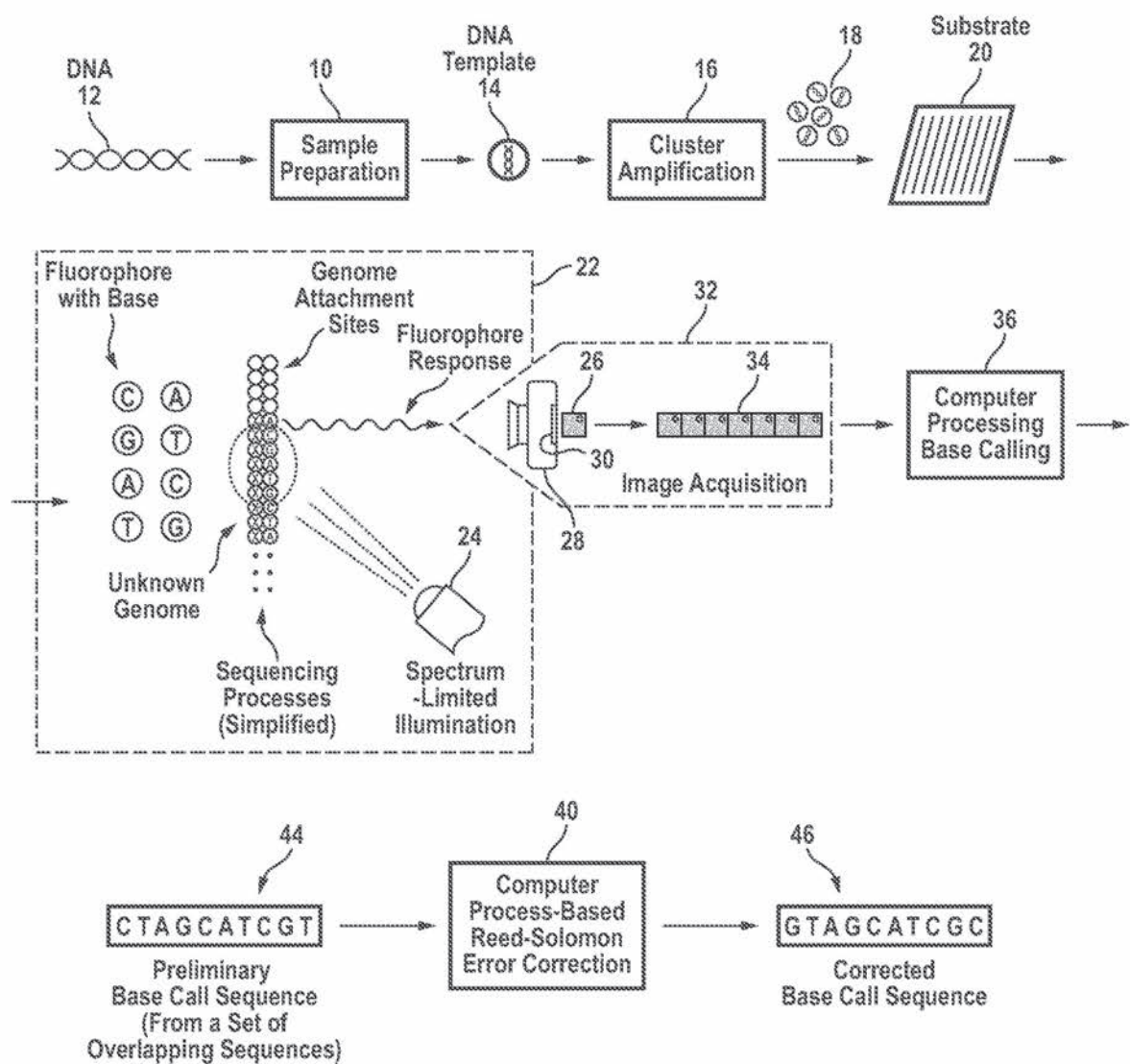


FIG. 1

Index	Observed Sequence	Corrected Sequence	Solomon Recovered	Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Score 7
2560	TACGTAG	TACGTAG	0	0.84	0.82	0.81	0.32	0.81	0.84	0.09
2561	TGGCTAA	TGGCAA	1	0.06	0.01	0.01	0	0.01	0.1	0.05
2562	TGGCAA	TGGCAA	0	0.94	0.86	0.87	0.87	0.85	0.9	0.95
2563	AGGAAAA	AGGAAAG	1	0	0.05	0.04	0	0	0.01	0
2564	AGGAAAG	AGGAAAG	0	0.13	0.21	0.28	0.12	0.21	0.05	0.12

FIG. 2

Categories	Observed Sequences	Reed-Solomon Corrected Sequences
Number of Valid Sequences	20647	35317
Number of UnCalled Sequences	0	15308
Number of Invalid Sequences	29978	0
Total	50625	50625

FIG. 3

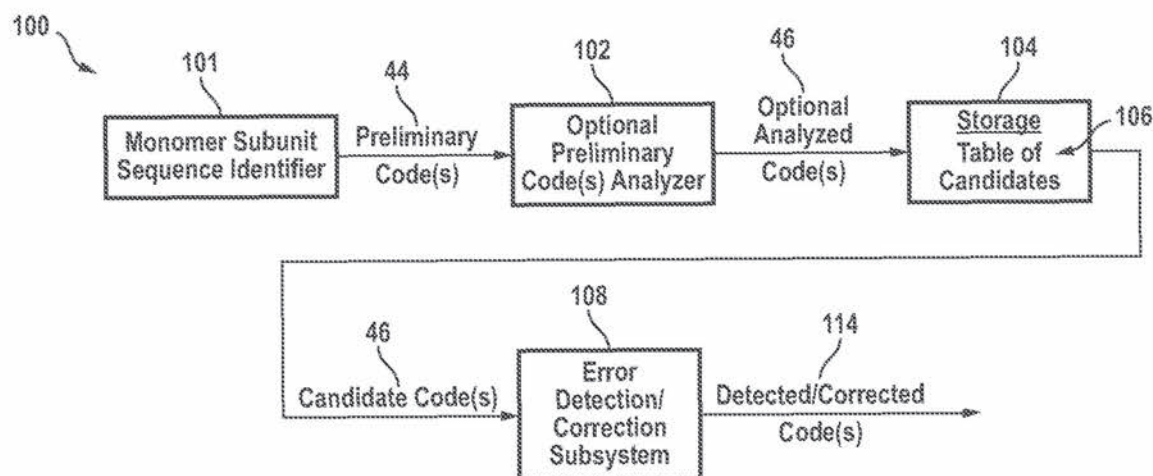


FIG. 5

Index	Observed Sequence	Corrected Sequence	Reed-Solomon Recovered	Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Score 7	Score 8	Score 9	Score 10
20	CGGTTCTTAT	CGGTACTTAT	1	0.1574	0.4825	0.2193	0.4314	0.0104	0.2906	0.2856	0.0262	0.0254	0.1353
29	GTAAGACTGC	GTAAGACTGC	0	0.9143	0.8828	0.9396	0.9122	0.8947	0.9499	0.959	0.9488	0.9444	0.9335
30	AGACTGGCGC	AGACTGGAGC	1	0.0968	0.0579	0.114	0.0141	0.0735	0.2902	0.3316	0.0029	0.4844	0.5554
37	AGGACACCGC	AGGATACCGC	1	0.9943	0.0615	0.5716	0.6492	0.0587	0.1396	0.1302	0.1057	0.1038	0.9983
40	CGCGAGTCGA	CGCGAGTCGA	0	0.1824	0.886	0.9227	0.911	0.9232	0.9505	0.9399	0.97	0.9421	0.9668
50	AAAAGTCCGG	AAAAGTCCGG	0	0.8932	0.8876	0.8989	0.8734	0.8854	0.2525	0.6954	0.8807	0.9075	0.9415
68	ACTTTATTAC	*****	0	0	0.0007	0.0406	0	0	0	0	0	0	0
69	ACTAACCGTA	ACTAACCGTA	0	0.8452	0.8646	0.8521	0.8821	0.6185	0.8762	0.882	0.8881	0.8917	0.882
70	CACATGTTAC	CACATGTTAC	0	0.9579	0.9647	0.981	0.974	0.9926	0.9801	0.9575	0.9429	0.9281	0.9401
71	TTGCGGGCTT	TTGCGGGCCT	1	0.4263	0.6358	0.3367	0.1561	0.4507	0.2606	0.2678	0.2049	0.0075	0.5003
73	TACTCAGTCA	CCTCAGTCA	1	0	0.3616	0.3828	0.6239	0.174	0.2373	0.0189	0.3441	0.0502	0.0959
74	CTAGAGTTTC	CTAGAGTTTC	0	0.3268	0.9163	0.9362	0.9205	0.9314	0.9724	0.9342	0.9254	0.9255	0.9675
85	ACACCACCGC	ACACCACCGC	0	0.7051	0.9319	0.903	0.9003	0.8817	0.9548	0.9731	0.9805	0.9563	0.9484
109	ACACAACGCC	ACACGACGCC	1	0.7707	0.0847	0.8321	0.0349	0.0093	0.9404	0.9251	0.9244	0.9463	0.9096

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 10/23083

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - H04L 1/00 (2010.01)

USPC -

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC- H04L 1/00 (2010.01)

USPC- 714/746

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
 PubWest (US Pat, PgPub, EPO, JPO), DialogClassic (Derwent, EPO, JPO, USPTO, WIPO), GoogleScholar: process, error, correct, sequence, nucleotide, nucleic, dna, reed, solomon, fluorophore, nanoball, nano particle

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X — Y	US 2007/0113137 A1 (RYU) 17 May 2007 (17.05.2007) para [0005]-[0021]	1, 5-7, 12, 17-18 2-4, 8-11, 13-16, 19-20
Y	US 4,584,686 A (FRITZE) 22 April 1986 (22.04.1986) col 2, ln 16-40; col 6, ln 15-68; col 13, ln 58 to col 14, ln 50; bol 19, ln 6-33	2-4, 13-16
Y	US 2008/0320358 A1 (PANDEL et al.) 25 December 2005 (25.12.2005) para [0016]-[0020], [0024]	8-11, 19-20
A,P	US 2009/0172501 A1 (LABLANS) 02 July 2009 (02.07.2009),	1-20
A	US 7,228,485 B1 (WU et al.) 05 June 2007 (05.06.2007)	1-20
A	US 6,430,723 B2 (KODAMA et al.) 06 August 2002 (06.08.2002)	1-20

☐ Further documents are listed in the continuation of Box C.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search
 13 March 2010 (13.03.2010)

Date of mailing of the international search report

30 MAR 2010

Name and mailing address of the ISA/US
 Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
 P.O. Box 1450, Alexandria, Virginia 22313-1450
 Facsimile No. 571-273-3201

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300
 PCT OSP: 571-272-7774