(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 October 2013 (24.10.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/159042 A1

(51) International Patent Classification:
G01N 27/64 (2006.01)

(21) International Application Number:
PCT/US2013/037454

(22) International Filing Date:
19 April 2013 (19.04.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/635,768 19 April 2012 (19.04.2012) US

(71) Applicant: **UNIVERSITY OF WASHINGTON THROUGH ITS CENTER FOR COMMERCIALIZATION** [US/US]; 4311 11th Avenue NE, Suite 500, Seattle, WA 98105-4608 (US).

(72) Inventors: **GUNDLACH, Jens, H.**; 13900 Northwood Road NW, Seattle, WA 98177 (US). **DERRINGTON, Ian, M.**; 3636 Dayton Avenue North, #4, Seattle, WA 98103 (US). **LASZLO, Andrew**; 6212 Pheasant Ct., Fort Collins, CO 80525 (US). **MANRAO, Elizabeth**; 13223 35th Avenue NE, #2, Seattle, WA 98125 (US).

(74) Agent: **NOWAK, Thomas, S.**; Christensen O'Connor Johnson Kindness PLLC, 1420 Fifth Avenue, Suite 2800, Seattle, WA 98101-2347 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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,

[Continued on next page]

(54) Title: METHODS AND COMPOSITIONS FOR GENERATING REFERENCE MAPS FOR NANOPORE-BASED POLYMER ANALYSIS

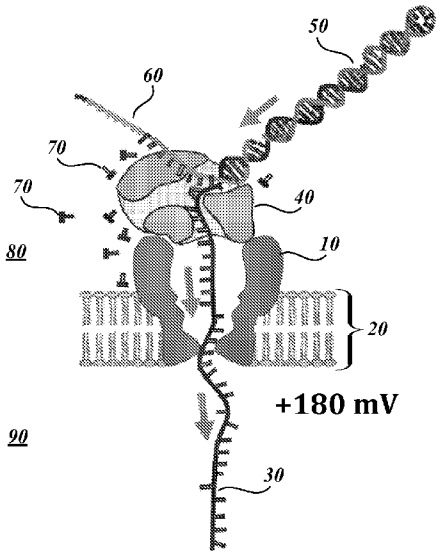

Fig.1A.

(57) Abstract: The present disclosure generally relates to the methods and compositions to efficiently analyze polymer characteristics using nanopore-based assays. Specifically disclosed is a method for generating reference signals for polymer analysis in a nanopore system, wherein the nanopore system has a multi-subunit output signal resolution. The method comprises translocating a reference sequence through a nanopore to generate a plurality of reference output signals, wherein each possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence. The output signals are compiled into a reference map for nanopore analysis of an analyte polymer. Also provided are methods and compositions for calibrating the nanopore system for optimized polymer analysis.

WO 2013/159042 A1

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

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

METHODS AND COMPOSITIONS FOR GENERATING REFERENCE MAPS FOR NANOPORE-BASED POLYMER ANALYSIS

CROSS-REFERENCE TO RELATED APPLICATION

This application claims the benefit of U.S. Application No. 61/635,768, filed April 19, 2012.

STATEMENT REGARDING SEQUENCE LISTING

The sequence listing associated with this application is provided in text format in lieu of a paper copy and is hereby incorporated by reference into the specification. The name of the text file containing the sequence listing is 40766_Seq_Final_2013-04-19.txt. The text file is 5 KB; was created on April 19, 2013; and is being submitted via EFS-Web with the filing of the specification.

STATEMENT OF GOVERNMENT LICENSE RIGHTS

This invention was made with Government support under NHGRI R01HG005115 and HG006321 awarded by the National Institutes of Health. The Government has certain rights in the invention.

BACKGROUND

The rapid, reliable, and cost-effective analysis of polymer molecules, such as sequencing of nucleic acids and polypeptides, is a major goal of researchers and medical practitioners. The ability to determine the sequence of polymers, such as a nucleic acid sequence in DNA or RNA, has additional importance in identifying genetic mutations and polymorphisms. Established DNA sequencing technologies have considerably improved in the past decade but still require substantial amounts of DNA and several lengthy steps and struggle to yield contiguous readlengths of greater than 100 nucleotides. This information must then be assembled "shotgun" style, an effort that depends non-linearly on the size of the genome and on the length of the fragments from which the full genome is constructed. These steps are expensive and time-consuming, especially when sequencing mammalian genomes.

Nanopore-based analysis methods have been investigated as an alternative to traditional polymer analysis approaches. These methods involve passing a polymeric molecule, for example single-stranded DNA ("ssDNA"), through a nanoscopic opening while monitoring a signal such as an electrical signal that is influenced by the physical properties of the target molecule as it passes through the nanopore opening. The

nanopore optimally has a size or three-dimensional configuration that allows the polymer to pass only in a sequential, single file order. Under theoretically optimal conditions, the polymer molecule passes through the nanopore at a rate such that the passage of each discrete monomeric subunit of the polymer can be correlated with the monitored signal. Differences in the chemical and physical properties of the monomeric subunits that make up the polymer, for example, the nucleotides that compose the ssDNA, result in characteristic electrical signals. However, nanopores that have been heretofore used for analysis of DNA and RNA, for example, protein nanopores held within lipid bilayer membranes and solid state nanopores, have generally not been capable of reading a sequence at a single-nucleotide resolution. Accordingly, the monitored signals must undergo a deconvolution step to deduce a correlation between the observed signal and the physical characteristics of the monomeric subunits passing through the nanopore. Furthermore, minor fluctuations in assay conditions or nanopore characteristics can differentially influence the monitored signals produced, thus making comparisons between assays, even ones using the same pore, difficult.

Accordingly, a need remains to efficiently correlate the observed signals from existing and future nanopore-based analysis systems to reliably ascertain the physical characteristics of the polymers applied thereto. The methods of the present disclosure addresses this and related needs of the art.

SUMMARY

This summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This summary is not intended to identify key features of the claimed subject matter, nor is it intended to be used as an aid in determining the scope of the claimed subject matter.

In one aspect, the present disclosure provides a method for generating reference signals for polymer analysis in a nanopore system wherein the nanopore system has a multi-subunit output signal resolution. The method comprises translocating a reference sequence through a nanopore of the nanopore system to generate a plurality of reference output signals, wherein each possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence. The method also comprises compiling the plurality of reference output signals into a reference map for nanopore analysis of an analyte polymer.

In some embodiments, the output signal is an ion current through the nanopore. In some embodiments, the ion current is determined by the identities of a plurality of contiguous polymer subunits disposed in a constriction region of the nanopore. In some embodiments, the plurality of contiguous polymer subunits that determine the ion current output signal comprise a dimer, trimer, quadromer, pentamer, hexamer, heptamer, octamer, nonamer or decamer of polymer subunits.

In some embodiments, the reference sequence is a De Bruijn sequence B represented by $B(k, n)$, wherein k is the number of potential polymer subunit identities, and wherein n is the number of contiguous polymer subunits that correspond to the resolution of the nanopore system. In some embodiments, the De Bruijn sequence is generated by taking a Hamiltonian path of an n-dimensional graph over k subunit identities, taking a Eulerian cycle of a $(n - 1)$-dimensional graph over k subunit identities, using finite fields analysis, or concatenating all possible Lyndon words whose length divides by n.

In some embodiments, a segment of the reference sequence or the entire reference sequence is a reference sequence domain of a reference polymer. In some embodiments, the reference sequence domain of the reference polymer consists of the entire reference sequence. In some embodiments, the reference sequence domains of a plurality of distinct reference polymers each consists of an exclusive segment of the reference sequence, wherein the aggregate of the plurality of reference sequence domains contains the entire reference sequence. In some embodiments, each of the plurality of reference polymers further comprises an overlap domain proximal to the reference sequence domain, wherein the overlap domain consists of a 1-50 subunit polymer sequence appearing in the reference domain of another reference polymer.

In some embodiments, the reference polymer further comprises a marker domain with a noncanonical polymer subunit. In some embodiments, the noncanonical polymer subunit is a nucleic acid subunit selected from the group consisting of uracil, 5' methylcytosine, 5' hydroxymethylcytosine, 5' formethylcytosine, 5' carboxycytosine b-glucosyl-5-hydroxy- methylcytosine, 8-oxoguanine, or an abasic lesion.

In some embodiments, the reference polymer further comprises a calibration domain with a polymer sequence predetermined to provide a pattern of multiple output signals. In some embodiments, the analyte polymer comprises the same calibration domain as the reference polymer.

In some embodiments, the analyte polymer is a nucleic acid and the reference sequence is a nucleic acid sequence. In some embodiments, the nucleic acid is single stranded DNA or double stranded DNA. In some embodiments, the DNA nucleic acid comprises or consists of the canonical adenine, thymine, guanine and/or cytosine subunits. In some embodiments, the nucleic acid is RNA. In some embodiments, the RNA nucleic acid comprises or consists of the canonical adenine, uracil, guanine and/or cytosine subunits.

The method of Claim 1, wherein the analyte polymer is a polypeptide and the reference sequence is a polypeptide sequence.

The method of Claim 1, further comprising comparing one or more output signals generated by the nanopore system using an analyte polymer against the reference library to determine a characteristic of the target polymer.

In some embodiments, the characteristic of the analyte polymer is a sequence pattern of the target polymer. In some embodiments, the characteristic of the analyte polymer is a primary sequence of the target polymer. In some embodiments, the reference sequence is translocated through the nanopore by an electrophoretic and/or an enzymatic mechanism. In some embodiments, the analyte polymer is translocated through a nanopore by the same mechanism as the reference sequence.

In another aspect, the disclosure provides a method for assessing the utility of a nanopore system for polymer analysis, wherein the nanopore system has a multi-subunit output signal resolution. The method comprises translocating a reference sequence through a nanopore of the nanopore system to generate a plurality of reference output signals, wherein each possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence. The method also comprises compiling the plurality of reference output signals into a reference map. The method also comprises determining the frequency of degenerate output signals, whereby a low level of degeneracy is indicative of a high utility of the nanopore system for polymer analysis.

DESCRIPTION OF THE DRAWINGS

The foregoing aspects and many of the attendant advantages of this invention will become more readily appreciated as the same become better understood by reference to the following detailed description, when taken in conjunction with the accompanying drawings, wherein:

FIGURE 1A illustrates a cross-sectional view of a representative nanopore 10 and related system components for polymer analysis. The illustrated system incorporates a protein nanopore 10 embedded in a lipid bilayer 20. A DNA polymerase 40 is bound to the DNA analyte, which contains a single stranded DNA (ssDNA) segment 30 and a double stranded DNA (dsDNA) segment 50. As illustrated, the ssDNA polymer 30 is translocating from the *cis* side 80 to the *trans* side 90 of the nanopore 10, as indicated by the directional arrows. As the ssDNA 30 translocates, a blocking oligomer 60 is unzipped (i.e., decomplemented) away from the ssDNA analyte 30 template by pull force exerted on the DNA analyte. Once the blocking oligomer 60 is completely unzipped from the ssDNA template, an extendible end of the double stranded DNA segment 50 is exposed and capable of extension by the DNA polymerase 40 (see FIGURE 1B below).

FIGURE 1B illustrates a cross-sectional view of the representative nanopore 10 and related system components illustrated in FIGURE1A. However, in this illustration, the ssDNA polymer 30 is translocating from the *trans* side 90 to the *cis* side 80 of the nanopore 10, as indicated by the directional arrows. In this stage, the polymerase activity of the DNA polymerase 40 pulls the ssDNA 30 through the nanopore 10 as it incorporates dNTPs 70 to extend the double stranded portion 50 of the analyte using the ssDNA 10 as the template.

FIGURE 2 illustrates a representative quadromer map generated from a reference DNA sequence described herein. The current measurements for all 256 possible nucleotide subunit quadromers are illustrated as a fraction of open pore current. These values require multiplication by ~115 pA to convert to pico ampere units.

DETAILED DESCRIPTION

The present disclosure generally relates to the methods to efficiently analyze polymer characteristics using nanopore-based assays.

Nanopores hold promise for inexpensive, fast, and nearly "reagent-free" analysis of polymers. In a general embodiment, an external voltage is applied across a nanometer-scale, electrolyte filled pore inducing an electric field. Any analyte, such as a polymer, that resides in, or moves through, the interior of the pore modulates the ionic current that passes through the pore depending on its physical characteristics. If the interior tunnel formed by the pore is of sufficiently small diameter and length, polymers that pass through must pass in a linear fashion, such that only a subset of the polymer subunits reside in the most constricted zone of the pore tunnel at one time. Thus, the ionic current

fluctuates over time as the polymer passes through the nanopore, subunit by subunit, depending on the different physical characteristics of the subunit(s) residing in the nanopore constriction zone at each iterative step.

Ideally, a polymer in a nanopore will generate a unique output signal, such as a current level, that is determined by a single, monomeric subunit of the polymer residing in the pore at each iterative translocation step. Thus, as the polymer translocates the resulting trace of output signals can be translated directly into the primary sequence of the polymer. However, most nanopores have constriction zones that physically interact with more than one polymer subunit during any single translocation event and thus the magnitude of the ion current is influenced by more than one subunit. Accordingly, the resulting output signal reflects a plurality of subunits. Each output signal then must be correlated to signals produced by known combinations of polymer subunits. Thus, for polymer analysis, a reference map is required to deduce the sequence or other physical characteristics of the analyte polymer.

The present inventors have developed an efficient approach to generating a reference map for any nanopore system capable of translocating a polymer. The approach, described herein, has the advantages of minimizing the length of the reference sequence required to generate a comprehensive reference map for any polymer where the subunits are known. This results in reduced costs and efforts in generating reference polymers, reduced translocation complications associated with overlong reference polymers, and wide applicability to various nanopore-types.

In accordance with the foregoing, in one aspect, the present disclosure provides a method for generating reference signals for polymer analysis in a nanopore system. The nanopore systems addressed by the present disclosure have a multi-subunit output signal resolution. The method comprises translocating a reference sequence through a nanopore of the nanopore system to generate a plurality of reference output signals, wherein every or substantially every possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence; and compiling the plurality of reference output signals into a reference map for nanopore analysis of an analyte polymer. In some embodiments, every possible multi-subunit sequence that can determine an output signal appears in the reference sequence. In some embodiments, every possible multi-subunit sequence that can determine an output signal appears in the reference sequence only once. In some embodiments, substantially every possible multi-subunit sequence that can

determine an output signal appears in the reference sequence. In some embodiments, substantially every possible multi-subunit sequence that can determine an output signal appears in the reference sequence only once.

Many aspects of the nanopore systems of the present disclosure are familiar in the art. Typically, a pore is disposed within a nonconductive barrier between a first conductive liquid medium and a second conductive liquid medium. The nanopore, thus, provides liquid communication between the first and second conductive liquid media. In some embodiments, the pore provides the only liquid communication between the first and second conductive liquid media. The liquid media typically comprises electrolytes or ions that can flow from the first conductive liquid medium to the second conductive liquid medium through interior of the nanopore. Additionally, the analyte polymer as the target or focus of an analysis is capable of entering the nanopore and translocating, preferably in a linear fashion, through the pore to the other side. In some cases, the first and second conductive liquid media located on either side of the nanopore are referred to as being on the *cis* and *trans* regions, where the analyte polymer to be measured generally translocates from the *cis* region to the *trans* region through the nanopore. However, in some embodiments, the analyte polymer to be measured can translocate from the *trans* region to the *cis* region through the nanopore. In some cases, the polymer as a whole does not pass through the pore, but portions or segments of the polymer pass through the nanopore for analysis.

Nanopores useful in the present disclosure include any nanopore capable of permitting the linear translocation of a polymer from one side to the other at a velocity amenable to monitoring techniques, such as techniques to detect current fluctuations. In some embodiments, the nanopore comprises a protein, such as alpha-hemolysin or MspA. Protein nanopores have the advantage that, as biomolecules, they self-assemble and are essentially identical to one another. In addition, it is possible to genetically engineer protein nanopores to confer desired attributes, such as substituting amino acid residues for amino acids with different charges, or to create a fusion protein (e.g., an exonuclease+alpha-hemolysin). Thus, the protein nanopores can be wild-type or can be modified. For example, for descriptions of modifications to MspA nanopores, see U.S. Pat. Pub. No. 2012/0055792. In some cases, the nanopore is disposed within a membrane, or lipid bilayer, which can separate the first and second conductive liquid media. In some embodiments, the nanopore can be a solid state nanopore. Solid state

nanopores can be produced as described in U.S. 7,258,838 and U.S. 7,504,058. Solid state nanopores have the advantage that they are more robust and stable. Furthermore, solid state nanopores can in some cases be multiplexed and batch fabricated in an efficient and cost-effective manner. Finally, they might be combined with micro-electronic fabrication technology. In some cases, the nanopore comprises a hybrid protein/solid state nanopore in which a nanopore protein is incorporated into a solid state nanopore.

The analyte polymer and/or reference sequence can be translocated through the nanopore using a variety of mechanisms. For example, the analyte polymer and/ or reference sequence can be electrophoretically translocated through the nanopore. Additionally or alternatively, nanopore systems can include a component that translocates a polymer through the nanopore enzymatically. For example, a molecular motor can be included to influence the translocation of polymers through the nanopore. A molecular motor can be useful for facilitating entry of a polymer into the nanopore and/or facilitating or modulating translocation of the polymer through the nanopore. Ideally, the translocation velocity, or an average translocation velocity is less than the translocation velocity that would occur without the molecular motor. In any embodiment herein, the molecular motor can be an enzyme, such as a polymerase, an exonuclease, or a Klenow fragment. In one example, described in more detail below and illustrated in FIGURES 1A-1B, a DNA polymerase such as phi29 can be used to facilitate movement in both directions. See also Cherf, G.M., et al., "Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision," *Nat. Biotechnol. 30*:344-348 (2012). It is preferred that the output signals generated from an analyte polymer were generated in a system where the analyte polymer was translocated using the same mechanism that was used to translocate the reference sequence through its nanopore, which can be the same or different nanopore.

Nanopore systems also incorporate structural elements to apply a voltage across the nanopore-bearing membrane or film. For example, the system can include a pair of drive electrodes that drive current through the nanopores. Additionally, the system can include one or more measurement electrodes that measure the current through the nanopore. These can be a patch-clamp amplifier or a data acquisition device. For example, nanopore systems can include an Axopatch-1B patch-clamp amplifier (Axon

Instruments) to apply voltage across the bilayer and measure the ionic current flowing through the nanopore.

The disclosed methods are useful for nanopore systems that have a multi-subunit output signal resolution. As used herein, the term "multi-subunit output signal resolution" refers to the characteristic of the nanopore where any output signal, e.g., a measurable current level through the pore, is determined or influenced by the physical characteristics of two or more polymer subunits residing therein. Typically, the polymer subunits that determine or influence the output signal reside at that time in the "constriction zone," or three-dimensional region in the interior of the pore with the narrowest diameter. Depending on the length of the constriction zone, the number of polymer subunits that influence the passage of electrolytes, and thus current output signal, can vary. The output signal produced by the nanopore system is any measurable signal that provides a multitude of distinct and reproducible signals depending on the physical characteristics of the polymer or polymer subunits. For example, the current level through the pore is an output signal that can vary depending on the particular polymer subunit(s) residing in the constriction zone of the nanopore. As the polymer passes in iterative steps (e.g., subunit by subunit), the current levels can vary to create a trace of multiple output signals.

As used herein, the term "polymer" refers to a chemical compound comprising two or more repeating structural units, referred to herein interchangeably as "subunits," "monomeric units," or "mers," where each subunit can be the same or different. Nonlimiting examples of polymers to be analyzed with the present methods include: nucleic acids, peptides, and proteins, as well as a variety of hydrocarbon polymers (e.g., polyethylene, polystyrene) and functionalized hydrocarbon polymers, wherein the backbone of the polymer comprises a carbon chain (e.g., polyvinyl chloride, polymethacrylates). Polymers include copolymers, block copolymers, and branched polymers such as star polymers and dendrimers. The term "nucleic acid" refers to a deoxyribonucleotide polymer (DNA) or ribonucleotide polymer (RNA) in either single- or double-stranded form. The structure of the canonical polymer subunits of DNA, for example, are commonly known and are referred to herein as adenine (A), guanine (G), cytosine (C), and thymine (T). As a group, these are generally referred to herein as nucleotides or nucleotide residues. For RNA, the 20 canonical polymer subunits are the same, except with uracil (U) instead of thymine (T). Similarly, the structures of the

canonical polymer subunits of polypeptides are known, and are referred to as a group herein as amino acids or amino acid residues.

As described above, the methods of the present disclosure are useful for nanopore systems with multi-subunit signal resolution. Thus, the output signal, such as the ion current through the nanopore, is determined or influenced by more than one polymer subunit in the nanopore constriction site at any given time. In some embodiments, the resolution or output signal is determined by two or more contiguous polymer subunits in the nanopore at any given time. In some embodiments, the two or more contiguous polymer subunits are in the constriction zone of the nanopore thereby influencing the output signal. Embodiments include nanopore systems where the output signal is determined by two, three, four , five, six, seven, eight, nine, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more contiguous polymer subunits residing in the constriction zone of the nanopore at the time the output signal is measured (e.g., at the time the current level is measured). It is noted that the multi-subunit polymer sections that determine the output signal in the nanopore system can be referred to as "–mers", such as 2-mer, 3-mer, 4-mer, 5-mer, 6-mer, 7-mer, 8-mer, 9-mer, 10-mer, etc. Alternatively, the multi-subunit polymer sections can be referred to as a dimer, trimer, quadromer, pentamer, hexamer, heptamer, octamer, nonamer, decamer, etc. As described, the multi-subunit polymer sections influence the output signal, such as current level, because when residing in the narrow constriction zone, the physical properties of each subunit impedes the current flow through the nanopore. Thus, the identities of each subunit has an influence on, and is reflected in, the overall current level at that time.

The disclosed method comprises translocating a reference sequence through a nanopore of the nanopore system to generate a plurality of reference output signals. The reference sequence can be translocated electrophoretically by applying voltage to the system. Alternatively, or simultaneously, the sequence can be pulled or pushed through the nanopore, for example, with the action of a molecular motor.

The reference sequence is optimally one where each possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence. Thus the reference sequence is optimized to be as short as possible while providing every potential multi-subunit sequence that corresponds with the resolution of the nanopore system. For example, a nanopore system that as a resolution of four subunits generates an output signal determined or influenced by a quadromer of the polymer subunits. Thus

the reference sequence for this nanopore will have every possible quadromer sequence in as short a sequence as possible. This provides a distinct and surprising advantage of reducing the efforts and costs expended on reference sequence synthesis and the time and effort for applying the reference sequences to the nanopore system.

An optimized reference sequence with each multi-subunit appearing only once can be designed by generating a De Bruijn sequence. A De Bruijn *B* sequence is represented by $B(k, n)$, wherein k is the number of potential polymer subunit identities, and wherein n is the number of contiguous polymer subunits that correspond to the resolution of the nanopore system. A De Bruijn sequence that is represented as a circular cyclic sequence (i.e., with no beginning or end) will have a length of k^n. The De Bruijn sequence can be generated by taking a Hamiltonian path of an n-dimensional graph over k subunit identities, taking a Eulerian cycle of a $(n - 1)$-dimensional graph over k subunit identities, using finite fields analysis, or concatenating all possible Lyndon words whose length divides by n. An exemplary De Bruijn sequence B(4,4) (k is four to reflect the four possible nucleotide residues of the DNA "alphabet"; n is four to reflect a quadromer polymer segment that determines the output signal for some MspA-based nanopores systems) is set forth herein as SEQ ID NO:1 in linear form with an arbitrary start and end point. However, it should be noted that linearization of a cyclic or circular De Bruijn sequence interrupts the contiguity of $(n-1)$ multi-subunit segments, depending on the length of the segments n. For instance, for a De Bruijn reference sequence of quadromers, a linearizing breakpoint in the circular sequence interrupts the contiguity of three of the quadromers. Accordingly, three additional polymer subunits must be added to one end of the linear sequence to meet the criteria of having every possible quadromer appearing once.

Under some conditions, some multi-subunit sequences are difficult to synthesize or otherwise provide in a reference sequence. Accordingly it is understood that in some embodiments, fewer than every possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence. For example, substantially every multi-subunit sequence appears in the reference sequence. In this context, "substantially every possible multi-subunit sequence" refers to at least 90% of every possible multi-subunit sequence that can determine an output signal, such as 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of every possible multi-subunit sequence that can determine an output signal appears in the reference sequence.

Additionally, under some conditions, it may be desirable to introduce some sequence redundancy into the reference sequence itself. Such conditions can include difficulties of generating clear output signals due to secondary structure or differential translocation speed associated with particular sub-sequences. However, to maintain the advantages of the present disclosure, such redundancy is minimized. Accordingly, in some embodiments, at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of every possible multi-subunit sequence appears only once in the reference sequence. Consequently, the present disclosure encompasses embodiments where 10%, 9%, 8, 7, 6, 5, 4, 3, 2, or 1% of every possible multi-subunit sequence can appear more than once in the reference sequence, such as appearing two, three, four, or more times in the reference sequence.

In some embodiments, the entire reference sequence, which may consist of one of each multi-subunit polymer segment that can determine a signal output, is a reference sequence domain of a reference polymer molecule or construct. In some embodiments, the entire reference sequence is a reference sequence domain of a single reference polymer molecule or construct. In alternative embodiments, the reference sequence is segmented into reference subsequences. The reference sub-sequences or segments are mutually exclusive, meaning that there are no multi-subunit polymer sub-sequences in common among the reference sub-sequences or segments. For example, a De Bruijn sequence was developed for DNA to generate a quadromer map (for nanopores with output signals determined by four nucleotides) and segmented into eight subsequences, set forth herein as SEQ ID NOS:2-9. In some further embodiments, the reference sub-sequences or segments can be reference domains in a plurality of distinct reference polymers, as provided in Table 1, below. Each of the plurality of distinct reference polymers contains a distinct reference domain. The aggregate of the reference domains can be aligned or reassembled to provide the entire reference sequence without redundancy or extraneous polymer subunits.

As described above, when a breakpoint is introduced into a De Bruijn sequence, additional n-1 subunits must be added at one end to preserve the contiguity of all multi-subunit segments. Accordingly, in some embodiments, each of the plurality of reference polymers further comprises an overlap domain proximal to the reference sequence domain, wherein the overlap domain consists of a 3-50 subunit polymer sequence appearing in the reference domain of another reference polymer. As used herein, the term

"proximal" refers to the position directly preceding or subsequent to the reference position. In terms of DNA sequence, for example, the overlap domain is immediately 5' or 3' to the reference domain without intervening nucleotide residues. This restores the contiguity of the fragmented multi-subunit segments and allows for signal trace overlap (and adjustment) of traces generated from distinct reference polymers. The overlap domain corresponds to a subunit polymer sub-sequence appearing in a different reference domain. In the context of DNA polymers, typically, an overlap domain added to the 3' end of a first reference domain will correspond to (e.g., repeat) the sub-sequence at the 5' end of a second reference domain. Alternatively, an overlap domain added to the 5' end of a first reference domain will correspond to (e.g., repeat) the subsequence at the 3' end of a second reference domain. This type of configuration is illustrated for a group of eight reference domains and corresponding overlap domains in Table 1.

In some embodiments, the one or more reference polymers further comprise a marker domain with a noncanonical polymer subunit. A noncanonical subunit is useful to provide an obvious output signal to indicate that the end of the reference domain (and possibly overlap domain) has passed through the nanopore. Regarding embodiments of nucleic acid polymers, illustrative and nonlimiting examples of noncanonical subunits include uracil (for DNA), 5-methylcytosine, 5-hydroxymethylcytosine, 5-formethylcytosine, 5-carboxycytosine b-glucosyl-5-hydroxy-methylcytosine, 8-oxoguanine, 2-amino-adenosine, 2-amino-deoxyadenosine, 2-thiothymidine, pyrrolo-pyrimidine, 2-thiocytidine, or an abasic lesion. An abasic lesion is a location along the deoxyribose backbone but lacking a base. Representative noncanonical peptide residues are known in the art, as set forth in, for example, Williams et al., *Mol. Cell. Biol. 9*:2574 (1989); Evans et al., *J. Amer. Chem. Soc. 112*:4011-4030 (1990); Pu et al., *J. Amer. Chem. Soc.* 56:1280-1283 (1991); Williams et al., *J. Amer. Chem. Soc. 113*:9276-9286 (1991); and all references cited therein. Exemplary noncanonical amino acids include, but are not limited to:

2-Aminoadipic acid, N-Ethylasparagine, 3-Aminoadipic acid, Hydroxylysin, β-alanine, β-Amino-propionic acid, allo-Hydroxylysine, 2-Aminobutyric acid, 3-Hydroxyproline, 4-Aminobutyric acid, piperidinic acid, 4-Hydroxyproline, 6-Aminocaproic acid, Isodesmosine, 2-Aminoheptanoic acid, allo-Isoleucine, 2-Aminoisobutyric acid, N-Methylglycine, sarcosine, 3-Aminoisobutyric acid, N-Methylisoleucine, 2-Aminopimelic acid, 6-N-Methyllysine, 2,4-Diaminobutyric acid,

N-Methylvaline, Desmosine, Norvaline, 2,2'-Diaminopimelic acid, Norleucine, 2,3-Diaminopropionic acid, Ornithine, N-Ethylglycine. Methods of incorporating noncanonical amino acids are well known in the art.

In some embodiments, the reference polymer further comprises a calibration domain. In some embodiments, the calibration domain comprises a polymer sequence predetermined to provide a recognizable pattern of multiple output signals. In some embodiments, the multiple output signals are highly distinct to facilitate the pattern being easily recognizable. Thus, each output signal can be mutually distinguished with a high degree of reliability. For example, a useful pattern can be one that provides a high-low-high or a low-high-low current fluctuation. In some embodiments, both the high and the low current levels of the multiple output signals are near the extreme ends of the spectrum of possible output signal produced by multi-subunit polymer segments. This is useful to calibrate a plurality of signal (e.g., current) traces from different polymers. Because the calibration sequence is the same in each polymer (reference or analyte), the sequence pattern will be the same for every polymer. However, various assay conditions might fluctuate from assay to assay, thus causing the pattern to drift or shift upward or downward or changing conductance causing a variation in the calibration constant. Such changing conditions can include temperature, applied voltage, salinity of the conductive media, nanopore shape, etc. Using a calibration domain in every polymer where the domain contains a common sequence allows for the normalization of all sequence traces to a standardized set of levels. Furthermore, the calibration sequence can also be included in the analyte DNA and allows for rescaling the ion currents (or other conditions) to the same conditions as the reference oligonucleotide constructs. An exemplary calibration sequence for DNA is CCTTCCAAAAAAAA (set forth herein as SEQ ID NO:10), set forth below. Exemplary DNA reference polymers for generating a quadromer reference map, where the De Bruijn sequence was segmented into eight fragments, and where the reference polymers contain reference, overlap, marker and calibration domains, are provided below in Table 1 and are also set forth as SEQ ID NOS:11-18.

In accordance with the disclosed method, a reference sequence (for example, as contained in one or more reference polymers, as described above) appropriate for the resolution of a nanopore system is applied to the nanopore. As described, nanopores generally do not produce output signals that achieve single polymer subunit resolution

(i.e., output signals determined exclusively by a single polymer subunit). Nanopores can be readily assayed to determine their level of multi-subunit resolution. For example, test polymers with mostly uniform subunits can be translocated through the nanopore, and the number of output signals affected by the presence of a single, distinct subunit can be determined. An exemplary test polymer is described in Manrao et al., "Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore," *PLoS ONE* 6(10):e25723 (2011). In such a manner, it can be determined for any nanopore / polymer combination how many polymer subunits determine the output signal at each measurement. Once the "word" length (i.e., multi-subunit length) is determined, an optimized reference sequence can then be generated that corresponds to this nanopore resolution that contains every possible multi-subunit word of that length only once. The reference sequence, for example as contained in a single or multiple reference polymers, can be examined with to the nanopore. The output signals are thus generated for every possible multi-subunit segment from the shortest possible reference sequence. The output signals are compiled into a reference map or library that correlates each output signal with the sequence of each possible multi-subunit polymer segment that can generate a discrete output signal. An exemplary quadromer DNA map is described below and illustrated in Figure 2.

The reference map or library can thereafter be used to determine characteristics of an unknown analyte polymer. Output signals obtained from an analyte polymer are compared to the reference map. It will be understood that the analyte polymer is the same type of polymer as the reference polymer or reference sequence. For example, a reference nucleic acid polymer is used to generate a reference table for use in analyzing an analyte nucleic acid polymer. In some embodiments, the output signals for the analyte polymer are generated in the same type of nanopore as used to generate the output signals used to generate a reference map. As described above, nanopores such as protein nanopores can be generated with a high degree of similarity such that different copies of the same type (e.g., MspA, or one particular mutant thereof) are essentially identical to each other. Thus, a reference library generated using one particular MspA nanopore in a system can be used to analyze the output signals for an analyte measured using another MspA nanopore. As described, slight differences in assay conditions can be normalized using calibration domains present in the different reference and analyte polymers. In some embodiments, the output signals for the analyte polymer are generated in the same

nanopore as the output signals from the reference sequence used to generate a reference map.

In some embodiments, the characteristics ascertained from an analyte polymer can be a sequence pattern that does not provide a primary sequence but merely provides a unique and recognizable pattern of subunit structures, such as a "fingerprint." In this regard, the term "fingerprint" is used to refer to sufficient structural or sequence data that can be used to determine whether two polymers are different or whether they are likely the same. Alternatively, the sequence pattern can be the complete or primary sequence or the primary sequence of a portion of the polymer.

In practice, a trace obtained for an analyte polymer can be generated in a nanopore system. Each output signal can be compared to the reference map. When a matching output signal is found, the particular reference multi-subunit sequence is determined to have resided in the nanopore at that specific time of measurement. As a series of sequences are correlated to a series of output signals in the analyte polymer trace, the characteristics of the polymer (e.g., the sequence) can be reconstructed. For some nanopore systems, there will likely be some degeneracy of the output signals, i.e., where two or more distinct multi-subunit polymer sequences result in the same or similar output signal within some margin of error. Thus, when an analyte output signal matches a degenerate reference signal, all possible multi-subunit polymer sequences that correspond to that signal must be considered for that position. An algorithm can be applied to determine which of the multiple possible multi-subunit polymer sequences is correct. For example, Hidden-Markov Models (HMM) are particularly suited for recovering sequence information for this system. See, e.g., Timp W, et al., 2012, *Biophys. J. 102*:L37-L39, incorporated herein by reference. For example, as a general illustration of an embodiment, all potential multi-subunit polymer sequences for a degenerate sequence are stored. A second, subsequent output signal sequence is assessed to provide one or more potential multi-subunit polymer sequences for the second output signal. All potential multi-subunit polymer sequences for the first output signal that are incompatible with the potential multi-subunit polymer sequences for the second output signal are discarded. This process is repeated for more sub-sequence output signals. Thus, even if the first measurement yielded many possible multi-subunit polymer sequences, it is likely that after several measurements there will only be one or a few possible sequences that are consistent with all the measurements.

In another aspect, the disclosure provides compositions useful for performing the methods described above. For example, in some embodiments, the present disclosure provides a reference polymer that comprises a reference domain. In some embodiments, the reference polymer further comprises at least one of a marker domain and a calibration domain. In some embodiments, the present disclosure provides a plurality of reference polymers each comprising a distinct reference domain, wherein the aggregate of the reference domains establish a reference sequence as described above. In some embodiments, the plurality of reference polymers each further comprise an overlap domain as described herein.

The use of the term "or" in the claims is used to mean "and/or" unless explicitly indicated to refer to alternatives only or the alternatives are mutually exclusive, although the disclosure supports a definition that refers to only alternatives and "and/or."

Following long-standing patent law, the words "a" and "an," when used in conjunction with the word "comprising" in the claims or specification, denotes one or more, unless specifically noted.

Disclosed are materials, compositions, and components that can be used for, can be used in conjunction with, can be used in preparation for, or are products of the disclosed methods and compositions. It is understood that, when combinations, subsets, interactions, groups, etc. of these materials are disclosed, each of various individual and collective combinations is specifically contemplated, even though specific reference to each and every single combination and permutation of these compounds may not be explicitly disclosed. This concept applies to all aspects of this disclosure including, but not limited to, steps in the described methods. Thus, if there are a variety of additional steps that can be performed, it is understood that each of these additional steps can be performed with any specific method steps or combination of method steps of the disclosed methods, and that each such combination or subset of combinations is specifically contemplated and should be considered disclosed. Additionally, it is understood that the embodiments described herein can be implemented using any suitable material such as those described elsewhere herein or as known in the art.

Publications cited herein and the material for which they are cited are hereby specifically incorporated by reference in their entireties.

The following is a description of an exemplary approach for generating a current signal in a nanopore-based system reflecting the sequence of target DNA polymers.

In nanopore-based polymer analysis, an ion current is measured as the polymer is passed or drawn through the opening of the nanopore. For DNA, it has been demonstrated that several nucleotides (i.e., an *n*-mer of DNA monomeric subunits) affect the ion current (see, e.g., Derrington et al., "Nanopore DNA sequencing with MspA," *PNAS 107*(37):16060-16065 (2010); Manrao et al., "Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore," *PLoS ONE 6*(10):e25723 (2011)). In the case of using a preferred nanopore, MspA mutant M2-NNN (MspA with the amino acid substitutions D90N, D91N, D93N, D118R, D134R, and E139K; Butler et al., "Single-molecule DNA detection with an engineered MspA protein nanopore," *PNAS 105*(52):20647-20652 (2008)), it was demonstrated that four nucleotides (i.e., nucleotide quadromers, or 4-mers) are involved in controlling the ion current when the DNA is held from the *cis* side and 180 mV is applied. When DNA is moved through the pore in single nucleotide steps, a succession of quadromer-generated current values best describes the ion current trace (Manrao et al., *Nature Biotechnology 30*:349-353 (2012)).

Briefly, as described in Manrao et al. (2012), the present inventors combined the previously engineered MspA nanopore (i.e., M2-NNN) with the phi29 DNA Polymerase (DNAP) blocking oligomer technique to read well-resolved and distinguishable current levels as DNA is drawn through the pore. In the assays, a single MspA pore was established in a lipid bilayer separating two chambers (*cis* and *trans*) containing 0.3 M KCl buffer solution. A patch-clamp amplifier applied +180 mV to the *trans* side of the bilayer and measured the ionic current through the pore. Current traces from the MspA-DNA-motor complex were analyzed in the unzipping phase and in the synthesis phase. Schematic illustrations of these phases are provided in FIGURES 1A and 1B, respectively.

For example, to study an easily resolvable DNA sequence, a 'block homopolymer' DNA template was used containing all four bases in short homopolymer sections of adenine (dA3), guanine (dG3) and cytosine (dC5), each separated by thymine (dT3). In control experiments without phi29 DNA polymerase (DNAP), the DNA translocation was too fast (>1 nt/μs) to read individual nucleotides. Next, phi29 DNAP was added to the *cis* volume, but the divalent cations and deoxyribonucleotide triphosphates (dNTPs) necessary for synthesis were omitted. Translocation events were recorded with current patterns consistent with force-activated unzipping of the blocking oligomer. These events exhibited distinct current levels including a high peak (~0.6 Io) when the two abasic sites

passed through the constriction. As expected, after removal of the blocking oligomer, the phi29 DNAP was unable to extend the primer strand owing to the absence of divalent cations and dNTPs. The phi29 DNAP eventually either fell off, allowing cooperative dissociation of the primer in MspA, or continued onwards, unzipping the primer strand. The duration of these events was generally >20 s.

To allow DNA synthesis to proceed, all four standard dNTPs (100 μM each) and 10 mM $MgCl_2$ were added. In five separate experiments using the block homopolymer DNA, 33 events with the same distinct current pattern were observed. A typical current trace was characterized by a rapid succession of current steps (levels) with various durations and current values ranging between 0.18 and 0.4 of I_o and two peaks with current levels above 0.6 I_o. Many levels at the beginning of an event were repeated later in the event, but in the opposite order. These results are consistent with part of each DNA molecule being read twice; once, during stepwise motion of the DNA template toward the *trans* side, while unzipping the blocking oligomer, and again, during stepwise motion back toward the *cis* side, while phi29 DNAP was synthesizing (see FIGURES 1A and 1B, respectively for an illustration of these phases). The levels observed during unzipping were symmetrical in time with levels observed during synthesis. Once phi29 DNAP completed synthesis of the DNA template strand, both the DNA and phi29 DNAP exited to the *cis* side. The average duration of events containing phi29 DNAP synthesis (~14 s) was shorter than events recorded in the absence of Mg^{2+} and dNTPs because synthesis is faster than the time for phi29 DNAP to release from the DNA template or to unzip the primer strand. In the experiments with $MgCl_2$ and dNTPs, >10% of the total data acquisition time was spent with a DNA-motor complex threaded through MspA.

The temporal ordering of the current levels was preserved across all traces and that the levels could be aligned with the known DNA sequence. The results were qualitatively consistent with the inventors' previous work (Manrao et al. (2011)) where homopolymer DNA was held statically in MspA by a NeutrAvidin molecule. The sequence of the DNA to the current levels were qualitatively aligned by assuming equal spacing of nucleotides and matching the position of the current peak associated with the abasic lesion, the turnaround point, and the low currents typically found with multiple thymines. The nucleotides positioned near the center of the constriction were found to dominate the total current of a level, whereas the nucleotides to either side control the

current to a lesser extent. This was consistent with previous MspA experiments where each current level was affected by ~four nucleotides in and around the constriction.

This and additional data indicated that a single-nucleotide influenced the current pattern for about four nucleotide steps through the nanopore. For example, a single-nucleotide substitution passing through MspA's constriction altered the current level pattern significantly, indicating that a nucleotide's identity and position in the strand was encoded in the observed current. This was consistent with the previously reported findings in Manrao et al. (2011) using a single-nucleotide polymorphism in NeutrAvidin-anchored DNA.

Also described in Manrao et al. (2012), the inventors demonstrated the feasibility of reading heterogeneous DNA with MspA and phi29 DNAP using four DNA heteromeric base sequences. Analogous to the block homopolymer DNA, these strands also had two abasic residues near the end of the template sequence to produce a clear marker signifying successful completion of synthesis. For each experimental sequence, consistent current patterns were obtained with distinct features corresponding to individual nucleotide steps as DNA passed through MspA's constriction. Current levels were extracted from 20 events (obtained from two different pores) that exhibited unzipping and synthesis action of the system. A consensus of a current level sequence was formed from multiple events. The events were aligned to the consensus using a Needleman-Wunsch algorithm and overlayed the aligned levels. Consistent with previous experiments, the current trace showed a symmetry about a level that corresponded to the nucleotides in MspA's constriction when the blocking oligomer was completely removed, demonstrating that current levels observed during unzipping are repeated in opposite time-order during synthesis. As the blocking oligomer used in this experiment was shorter (15 nt) than the length of the template to be read, the number of levels observed during unzipping was fewer than those observed during synthesis, and the abasic peak marking the end of the read was not observed during unzipping.

Ultimately, the inventors demonstrated that individual single-stranded DNA molecules traversing through the short and narrow constriction of MspA under the control of phi29 DNAP yielded distinct current levels that were related to the sequence of the nucleotides in MspA's constriction. Current patterns were associated with two distinct processes: the 5′ leading–motion of the DNA template, which is consistent with the nearly monotonic unzipping of the blocking oligomer, and the 3′ leading–motion, which

is consistent with synthesis by phi29 DNAP. Motion during synthesis was faster than unzipping, and the levels associated with most single-nucleotide advances could be identified. The DNA nucleotides that passed twice through MspA's constriction—once during unzipping of the blocking oligomer and once during synthesis—yielded redundant current readings. Thus, the inventors have shown that the high nucleotide sensitivity of MspA combined with the translocation control of phi29 DNAP enables single-nucleotide discrimination of DNA passing through a nanopore.

The following is a description of an exemplary approach for generating a quadromer reference map for nanopore-based analysis of DNA.

For sequence analysis of unknown DNA (or other polymers), the patterns of current levels obtained from nanopore-based analysis systems, such as those described above for DNA, need to be related to a known sequence. As described, the observed current signals for each iterative movement of a nucleotide subunit were determined by a plurality of nucleotides residing at that time in the pore constriction zone. Thus, a model predicting the current pattern will be complex, and deconvolution of a current trace to extract the underlying sequence will require comparison to a library of current levels for all possible combinations of the plurality of sequences, such as 4-mer sequences in the system described above.

To efficiently generate such a reference library, the present inventors measured the 256 current values for all the possible DNA quadromers, thus resulting in the Quadromer Map (QM). To optimize generation of all possible DNA quadromer current signals, the inventors designed reference DNA molecules that contained all 256 possible quadromer patterns (=4^4 because there are four possible monomeric subunits at each of four positions in the quadromer) in the shortest possible DNA sequence. Specifically, a De Bruijn sequence was generated, resulting in a circular, or cyclical sequence with the shortest possible sequence because each quadromer appears only once. Briefly, the De Bruijn sequence B is represented by $B(k, n)$, wherein k is the number of potential polymer subunit identities (i.e., size of the subunit alphabet) and n is the number of contiguous subunits reflecting, corresponding to, or determining the resolution of the nanopore system (i.e., length of the "word"). See, van Aardenne-Ehrenfest, T and de Bruijn, NG, 1951, "Circuits and trees in oriented linear graphs", Simon Stevin 28:203–217. This results in a circular reference sequence of only k^n characters long. For DNA and RNA molecules, the number of potential canonical polymer subunits (i.e., size of the alphabet,

or k) is 4. In the MspA nanopore systems described above the number of subunits that contribute to any one signal (i.e., length of the word, or n) is also 4. Various software packages exist to generate De Bruijn sequences. Of particular use was the De Bruijn Sequence Generator, by W. Owen Brimijoin, available at Matlab® Central website provided on the world wide web by The Mathworks, Inc. Thus, a De Bruijn sequence results in a reference polynucleotide molecule that is k^n characters, i.e., 256 nucleotides, long. In a linearized form, the De Bruijn sequence results in a reference sequence that is only 259 (i.e., k^n+n-1, or 256+4-1) nucleotides long and contains all possible quadromers in a contiguous orientation. This is a major advance in efficiency compared to the standard approach of concatamerizing each potential nucleotide quadromer, which results in a reference polynucleotide of 1024 nucleotides (i.e., total possible sequences each with four nucleotides, represented by k^n x 4).

With this approach, a circular or cyclical De Bruijn sequence was generated, as set forth herein as SEQ ID NO:1 with arbitrary start and end points. Because any oligonucleotide with the reference sequence must be drawn through the nanopore using the same mechanism as the analyte DNA polymer(s), the circular De Bruijn sequence design was linearized. Additionally, the linear reference sequence design with the De Bruijn sequence was separated into eight distinct segments, which were used as the design templates for synthesis of the reference oligonucleotide constructs. This segmentation was performed to further facilitate the translocation of the resulting reference oligonucleotide constructs through the MspA nanopore. The eight reference sequence segments, set forth herein as SEQ ID NOS:2-9, contain unique and exclusive portions of the full length reference sequence. These sequences are listed in the "Reference Domain" column of Table 1 in descending order.

Furthermore, each reference oligonucleotide construct was generated with an additional overlap sequence to assist continuity between the signals generated from each reference oligonucleotide construct and to provide that every possible quadromer was presented in a contiguous sequence. Specifically, each reference oligonucleotide construct also contain an additional five nucleotide sequence proximal to the 3' end of the reference sequence domain. The additional nucleotide sequence duplicated (or "overlapped with") the first five nucleotide sequence at the 5' end of the subsequent reference sequence segment, which appeared in a different reference oligonucleotide construct. These sequences are listed in the "Overlap Domain" column of Table 1.

Proximal the overlap domain sequence, each reference oligonucleotide construct contained an abasic site (i.e., a location or position along the deoxyribose backbone lacking a base). The abasic site, listed as an "N" in the "Marker Domain" column in Table 1, served to provide a clear and unambiguous reference point in the current trace generated by the passage of this nucleotide position through the nanopore.

Table 1: Sequences for the domains of eight reference DNA polymers that provide one De Bruijn reference sequence. The sequences of the domains are presented in the order as they appear in the full length sequence from the 5' end to the 3' end. The designations for the full-length sequences are provided in the right-hand column.

Reference Domain	Overlap Domain	Marker Domain	Calibration Domain	SEQ ID NO: for Full Sequence
CTTTTCTTCCTCTCCCCAAAAGAAATAAACAA	GGAAG	N	CCTTCCAAAAAAAA	11
GGAAGTAAGCAATGAATTAATCAACGAACTAA	CCAGA	N	CCTTCCAAAAAAAA	12
CCAGAGATAGACAGGGAGGTAGGCAGTGAGTT	AGTCA	N	CCTTCCAAAAAAAA	13
AGTCAGCGAGCTAGCCATATACATGGATGTAT	GCATT	N	CCTTCCAAAAAAAA	14
GCATTGATTTATTCATCGATCTATCCACACGG	ACGTA	N	CCTTCCAAAAAAAA	15
ACGTACGCACTGACTTACTCACCGACCTACCC	GGGGT	N	CCTTCCAAAAAAAA	16
GGGGTGGGCGGTTGGTCGGCTGGCCGTGTGCG	TTTGT	N	CCTTCCAAAAAAAA	17
TTTGTTCGTCTGTCCGCGCTTGCTCGCCTGCC	CTTTT	N	CCTTCCAAAAAAAA	18

Finally, each reference oligonucleotide construct contained a calibration sequence proximal to the marker domain and at the 3' end of full-length sequence. The calibration sequence, CCTTCCAAAAAAAA (set forth herein as SEQ ID NO:10) served to provide a recognizable pattern of high and low current patterns for each signal output trace. Because the recognizable pattern is known to possess the same sequence in every reference oligonucleotide, the remaining output signals in each trace can be calibrated to, or adjusted to match, the traces obtained from the other reference oligonucleotides. This calibration allows for normalization of signal variation obtained from different oligonucleotides that are due to variations in the assay conditions, such as salinity, temperature, viscosity, or the shape of the nanopore construction region. Furthermore,

the calibration sequence can also be included in the analyte DNA and allows for rescaling the ion currents (or other conditions) to the same conditions as the reference oligonucleotide constructs.

While Table 1 separately lists the sequences of the distinct Reference, Overlap, Marker, and Calibration domains of the eight reference oligonucleotide constructs (i.e., reference polymers) in their respective domain columns, the complete sequences for the reference oligonucleotide constructs are set forth herein as SEQ ID NOS:11-18, as indicated in the right-hand column.

While illustrative embodiments have been illustrated and described, it will be appreciated that various changes can be made therein without departing from the spirit and scope of the invention.

CLAIMS

The embodiments of the invention in which an exclusive property or privilege is claimed are defined as follows:

1. A method for generating reference signals for polymer analysis in a nanopore system, wherein the nanopore system has a multi-subunit output signal resolution, the method comprising:

translocating a reference sequence through a nanopore of the nanopore system to generate a plurality of reference output signals, wherein each possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence; and

compiling the plurality of reference output signals into a reference map for nanopore analysis of an analyte polymer.

2. The method of Claim 1, wherein the output signal is an ion current through the nanopore.

3. The method of Claim 2, wherein the ion current is determined by the identities of a plurality of contiguous polymer subunits disposed in a constriction region of the nanopore.

4. The method of Claim 3, wherein the plurality of contiguous polymer subunits that determine the ion current output signal comprise a dimer, trimer, quadromer, pentamer, hexamer, heptamer, octamer, nonamer or decamer of polymer subunits.

5. The method of Claim 1, wherein the reference sequence is a De Bruijn sequence B represented by $B(k, n)$, wherein k is the number of potential polymer subunit identities, and wherein n is the number of contiguous polymer subunits that correspond to the resolution of the nanopore system.

6. The method of Claim 5, wherein the De Bruijn sequence is generated by taking a Hamiltonian path of an n-dimensional graph over k subunit identities, taking a Eulerian cycle of a $(n - 1)$-dimensional graph over k subunit identities, using finite fields analysis, or concatenating all possible Lyndon words whose length divides by n.

7. The method of Claim 1, wherein a segment of the reference sequence or the entire reference sequence is a reference sequence domain of a reference polymer.

8. The method of Claim 7, wherein the reference sequence domain of the reference polymer consists of the entire reference sequence.

9. The method of Claim 7, wherein the reference sequence domains of a plurality of distinct reference polymers each consists of an exclusive segment of the reference sequence, wherein the aggregate of the plurality of reference sequence domains contains the entire reference sequence.

10. The method of Claim 9, wherein each of the plurality of reference polymers further comprises an overlap domain proximal to the reference sequence domain, wherein the overlap domain consists of a 1-50 subunit polymer sequence appearing in the reference domain of another reference polymer.

11. The method of Claim 7, wherein the reference polymer further comprises a marker domain with a noncanonical polymer subunit.

12. The method of Claim 11, where the noncanonical polymer subunit is a nucleic acid subunit selected from the group consisting of uracil, 5' methylcytosine, 5' hydroxymethylcytosine, 5' formethylcytosine, 5' carboxycytosine b-glucosyl-5-hydroxy- methylcytosine, 8-oxoguanine, or an abasic lesion.

13. The method of Claim 7, wherein the reference polymer further comprises a calibration domain with a polymer sequence predetermined to provide a pattern of multiple output signals.

14. The method of Claim 13, wherein the analyte polymer comprises the same calibration domain as the reference polymer.

15. The method of Claim 1, wherein the analyte polymer is a nucleic acid and the reference sequence is a nucleic acid sequence.

16. The method of Claim 15, wherein the nucleic acid is single stranded DNA or double stranded DNA.

17. The method of Claim 15, wherein the nucleic acid is RNA.

18. The method of Claim 1, wherein the analyte polymer is a polypeptide and the reference sequence is a polypeptide sequence.

19. The method of Claim 1, further comprising comparing one or more output signals generated by the nanopore system using an analyte polymer against the reference library to determine a characteristic of the target polymer.

20. The method of Claim 19, wherein the characteristic of the analyte polymer is a sequence pattern of the target polymer.

21. The method of Claim 19, wherein the characteristic of the analyte polymer is a primary sequence of the target polymer.

22. The method of Claim 1, wherein the reference sequence is translocated through the nanopore by an electrophoretic and/or an enzymatic mechanism.

23. The method of Claim 22, wherein the analyte polymer is translocated through a nanopore by the same mechanism as the reference sequence.

24. A method for assessing the utility of a nanopore system for polymer analysis, wherein the nanopore system has a multi-subunit output signal resolution, the method comprising:

translocating a reference sequence through a nanopore of the nanopore system to generate a plurality of reference output signals, wherein each possible multi-subunit sequence that can determine an output signal appears only once in the reference sequence;

compiling the plurality of reference output signals into a reference map;

and determining the frequency of degenerate output signals, whereby a low level of degeneracy is indicative of a high utility of the nanopore system for polymer analysis.

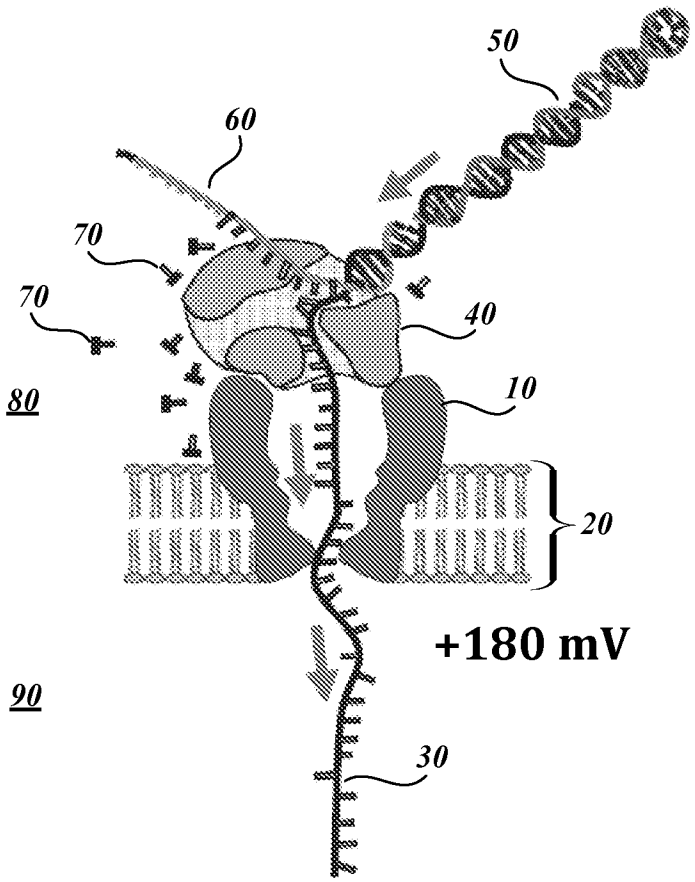

Fig.1A.

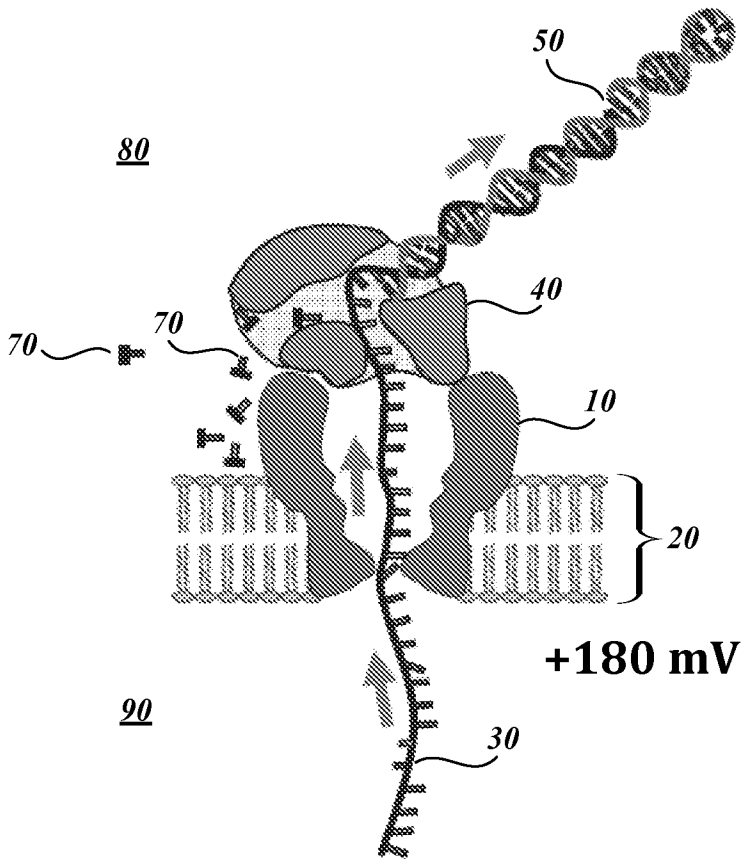

Fig.1B.

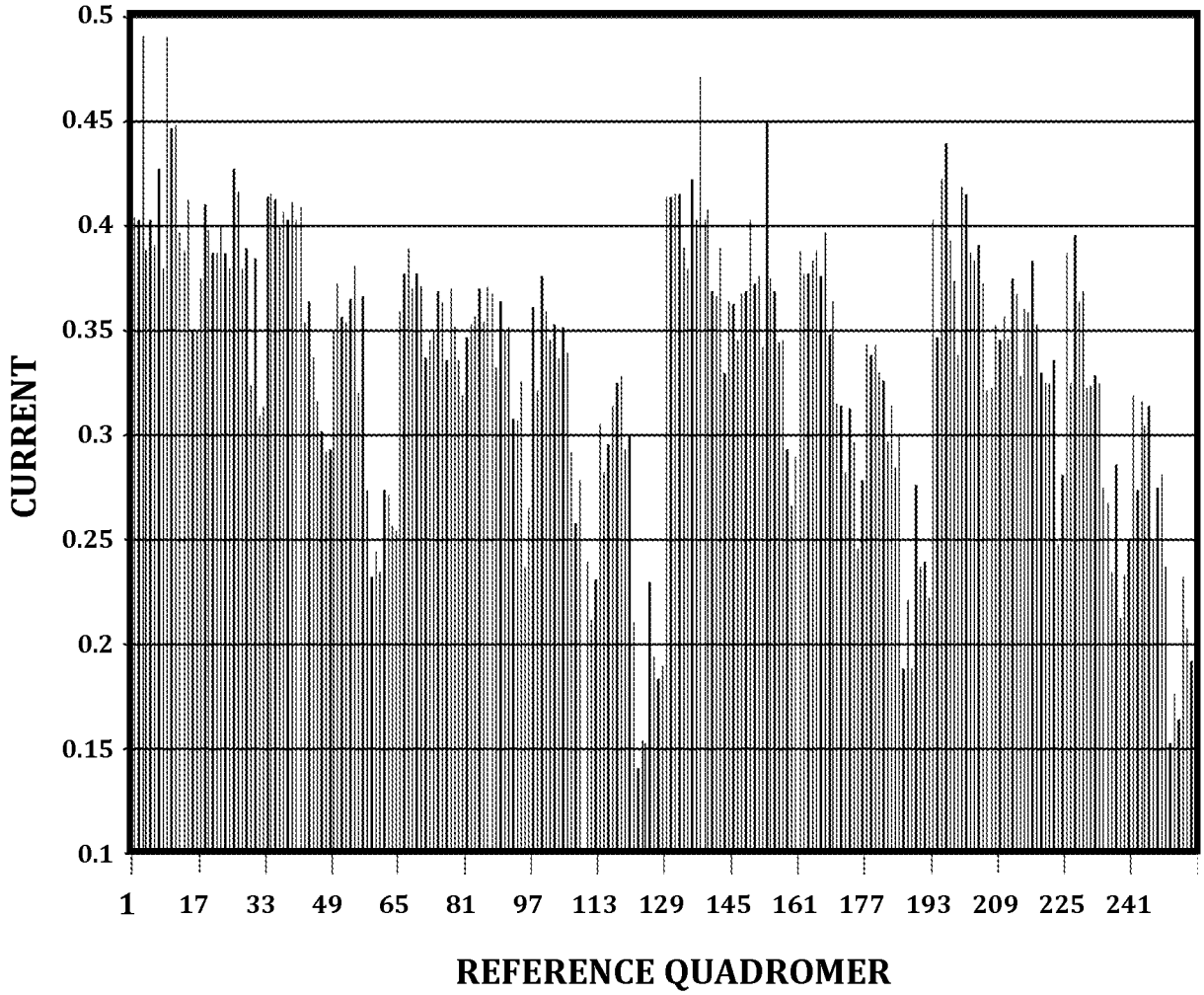
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
CURRENT
1
17
33
49
65
81
97
113
129
145
161
177
193
209
225
241
REFERENCE QUADROMER

Fig.2.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 2013/037454

A. CLASSIFICATION OF SUBJECT MATTER

G01N 27/64 *(2006.01)*

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)

C40B 20/00, C07K 14/705, C12N 9/12, C12Q 1/68, G01N 15/10, 15/12, 27/64, 27/66, 27/84, 33/68, G04C 3/04, 3/06

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)

PatSearch (RUPTO internal), USPTO, PAJ, Esp@cenet, Information Retrieval System of FIPS (http://www.fips.ru)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2008/124107 A1 (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA et al.) 16.10.2008	1-24
A	US 2003/0099951 A1 (MARK AKESON et al.) 29.05.2003	1-24
A	EP 1956367 A1 (PRESIDENT AND FELLOWS OF HARVARD COLLEGE et al.) 13.08.2008	1-24

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

☐ See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 July 2013 (08.07.2013)	15 August 2013 (15.08.2013)
Name and mailing address of the ISA/ FIPS Russia, 123995, Moscow, G-59, GSP-5, Berezhkovskaya nab., 30-1 Facsimile No. +7 (499) 243-33-37	Authorized officer I. Zhestovskaya Telephone No. (499) 240-25-91