POLYNUCLEOTIDE MAPPING AND SEQUENCING

RELATED APPLICATIONS

[0001] The present application claims the benefit of United States Application No. 61/115,704, filed November 18, 2008, the entirety of which is incorporated herein by reference for any and all purposes.

TECHNICAL FIELD

[0002] The disclosed invention relates to the field of nucleic acid sequencing and to the field of molecular imaging. The disclosed invention also relates to the field of nanotechnology.

BACKGROUND

[0003] With advances in molecular biology techniques has come increased interest in analyzing smaller and smaller samples with ever-increasing resolution and precision. Some of this is driven by the realization that population heterogeneity can often obscure salient features of a sample. Limited sample volume is also a consideration for some applications.

[0004] While existing techniques are, in theory, capable of extracting significant information from physically small samples, the effectiveness of such techniques has been limited by their ability to resolve structural features on such small samples. Accordingly, there is a need in the art for methods and related devices capable of obtaining genomic information based on single molecules or other physically small samples. The value of such methods would be enhanced if such methods were capable of improving upon the 1000 bp (1 kb) accuracy achieved by current techniques.

SUMMARY

[0005] In meeting the described challenges, the present disclosure first provides methods for assaying for the presence or relative positions of one or more exons, the methods comprising labeling first and second locations on a biopolymer with, respectively, a first and a second label such that the first and second labels flank a first region of the biopolymer that includes at least one constant exon; and linearizing the biopolymer and correlating the distance between the first and second labels to the presence, absence, or relative position of an alternative exon in said first region of the biopolymer.

[0005a] In an aspect, there is provided a method of assaying for the relative position of one or more exons, comprising:

labeling a biopolymer sample at a plurality of locations with a plurality of labels, wherein said labels are separated from one another by 30 base pairs to 1000 pairs on said biopolymer;

said plurality of labels including at least a first and second label located, respectively, at first and second locations on a biopolymer sample, such that the first and second labels flank a first region of the biopolymer sample that includes at least one constant exon;

determining the distance between the first and second labels; and

correlating the distance between the first and second labels to the presence, absence or relative position of an alternative exon in said first region of the biopolymer.

[0006] In an embodiment, the present disclosure provides methods of obtaining structural information about a DNA sample, comprising nicking a first double-stranded DNA

sample with a sequence-specific nicking endonucleoase; incorporating one or more dye-labeled nucleotides at two or more nicking sites effected by the nicking endonuclease; linearizing a portion of the first double-stranded DNA sample that includes at least two dye-labeled nucleotides; and registering the relative positions of two or more labeled dye-labeled nucleotides.

[0007] Also provided are methods of obtaining sequence information about a nucleic acid biopolymer, comprising binding a first fluorescently labeled sequence specific probe having a first binding sequence to a single-stranded nucleic acid biopolymer; contacting the single-stranded nucleic acid biopolymer with a first terminator nucleotide bearing a first fluorescent label, with a second terminator nucleotide bearing a second fluorescent label, with a third terminator nucleotide bearing a third fluorescent label, and with a fourth terminator nucleotide bearing a fourth fluorescent label; and linearizing and illuminating the nucleic acid biopolymer so as to determine the presence or relative positions of the first terminator nucleotide, the second terminator nucleotide, the third terminator nucleotide, the fourth terminator nucleotide, or any combination thereof, adjacent to the first labeled sequence-specific probe.

[0008] The present disclosure also provides methods of obtaining structural information about a nucleic acid biopolymer, comprising contacting a double-stranded biopolymer with a nicking endonuclease so as to effect a first nicking site; contacting the first nicking site with a first terminator nucleotide bearing a fluorescent label A, with a second terminator nucleotide bearing a fluorescent label B, with a third terminator nucleotide bearing a fluorescent label C, and with a fourth terminator nucleotide bearing a fluorescent label D; and linearizing and illuminating the double-stranded biopolymer so as to determine the relative positions of the first terminator nucleotide, the second terminator nucleotide, the third terminator nucleotide, the fourth terminator nucleotide, or any combination thereof.

[0009] Further provided are kits for performing multiplex hybridization, comprising a plurality of hybridization probes each of a different color; and instructions for applying at least two of the plurality of hybridization probes to a nucleic acid sample and linearizing and imaging at least one of the hybridized nucleic acids.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] The summary, as well as the following detailed description, is further understood when read in conjunction with the appended drawings. For the purpose of illustrating the invention, there are shown in the drawings exemplary embodiments of the invention; however, the invention is not limited to the specific methods, compositions, and devices disclosed. In addition, the drawings are not necessarily drawn to scale. In the drawings:

[0011] **Figure 1A** illustrates mapping statistics for Nt.BstNBI nicking endonuclease demonstrating that 100bp optical resolution dramatically improves the map accuracy and coverage;

[0012] **Figure 1B** illustrates unique mapping statistics for Nt.BspQI nicking endonuclease, demonstrating that 100bp optical resolution has little impact on the map accuracy and coverage;

[0013] **Figure 1C** illustrates that a comparatively fine map (1.5 kb) has better detection power for structural variations than does a comparatively coarse map (16kb);

[0014] **Figure 2A** depicts MAPT gene structure;

[0015] **Figure 2B** lists the size of each exon (alternative exons shown as shaded) of each exon present in the MAPT gene;

[0016] **Figure 2C** illustrates a barcoding or mapping scheme for super resolution imaging as applied to RNA exon splicing;

[0017] **Figure 2D** illustrates a multiplexed barcoding scheme;

[0018] **Figure 3** illustrates starting materials for sequencing;

[0019] **Figure 4** depicts the first cycle of a sequencing reaction;

[0020] **Figure 5** depicts the second sequencing cycle begun in **Figure 4**;

[0021] **Figure 6** demonstrates that a multiplexed sequencing scheme dramatically increases throughput;

[0022] **Figure 7A** depicts a model system of 741 bp PCR product used to demonstrate the resolution of SHRIMP;

[0023] **Figure 7B** illustrates imaging results after labeled DNA molecules were linearized on glass surface, indicating three (3) Cy3 dye molecules 30 nm and 60 nm apart, which was in good agreement with the 94bp and 172 bp distances between the three (3) Cy3 probes.

[0024] **Figure 8A** depicts a model system of a 741 bp PCR product used to demonstrate the resolution of SHRIMP and SHREC;

[0025] **Figure 8B** illustrates the imaging results after labeled DNA molecules were linearized on glass surface – the distances between Cy3-Cy5 pairs was 37±5 nm (32 nm expected) and 91±5 nm (87 nm expected), and the distance between Cy3-Cy3 pair to be 56±3 nm (58 nm expected) (**Figure 4**), demonstrating excellent agreement;

[0026] **Figure 9** depicts a sample, nonlimiting embodiment of the claimed methods of ascertaining structural information regarding genetic material;

[0027] **Figure 10** depicts a second sample, non-limiting embodiment of the claimed methods of ascertaining structural information regarding genetic material;

[0028] **Figure 11** depicts a non-limiting embodiment of the claimed methods; and

[0029] **Figure 12** depicts a further, non-limiting embodiment of the claimed methods.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

[0030] The present invention may be understood more readily by reference to the following detailed description taken in connection with the accompanying figures and examples, which form a part of this disclosure. It is to be understood that this invention is not limited to the specific devices, methods, applications, conditions or parameters described and/or shown herein, and that the terminology used herein is for the purpose of describing particular embodiments by way of example only and is not intended to be limiting of the claimed invention. Also, as used in the specification including the appended claims, the singular forms "a," "an," and "the" include the plural, and reference to a particular numerical value includes at least that particular value, unless the context clearly dictates otherwise. The term "plurality", as used herein, means more than one. When a range of values is expressed, another embodiment includes from the one particular value and/or to the other particular value. Similarly, when values are expressed as approximations, by use of the antecedent "about," it will be understood that the particular value forms another embodiment. All ranges are inclusive and combinable.

[0031] It is to be appreciated that certain features of the invention which are, for clarity, described herein in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention that are, for brevity, described in the context of a single embodiment, may also be provided separately or in any subcombination. Further, reference to values stated in ranges include each and every value within that range.

[0032] In a first embodiment, the present invention provides methods of assaying for the presence or even relative positions of one or more exons. These methods suitably include labeling first and second locations on a biopolymer sample with, respectively, a first and a second label such that the first and second labels flank a first region of the biopolymer sample that includes at least one constant exon. The user then correlates the distance between the first and second labels to the presence or absence (or relative position) of an alternative exon (i.e., an exon that does not appear in every mRNA) in said first region of the biopolymer. (The biopolymer is suitably DNA that is complementary to an mRNA; such DNA is easily synthesized by those of ordinary skill in the art.)

[0033] In some embodiments, the first and second labels are the same fluorophore. A wide range of fluorophores are suitable for the present incluvention, including the Cy- family of fluorophores. Other fluorophores will be known to those of skill in the art; a listing of fluorophores is found at, *e.g.*, http://info.med.yale.edu/genetics/ward/tavi/FISHdyes2.html. The labels may be of the same fluorophore, but may also be of different fluororphores.

[0034] The user suitably correlates the distance between the first and second labels to the presence, absence, or both of one or more alternative exons (or even the exons' relative positions) comprises comparing the distance between the first and second labels present on the biopolymer sample to the distance between labels that flank the first region of the biopolymer known not to contain an alternative axon. This is suitably accomplished by linearizing that region of the biopolymer that includes the fluorescent labels. Linearizing of biopolymers is discussed in detail in United States Patent Application No. 10/484,293 (granted Nov. 9, 2009), the entirety of which is incorporated herein by reference for all purposes.

[0035] The correlating suitably includes comparing the distance between the first and second labels to the distance between the first and second locations on a biopolymer lacking an alternative exon between the first and second locations.

[0036] This is illustrated by, e.g., Figure 2C, which figure depicts at "zero" a biopolymer having no alternative exons. Embodiment "2" in that figure depicts a biopolymer having alternative exon "2", which alternative exon may be detected by observing that the exon results in an increased separation distance (342 bp) between the Cy3 and Cy5 dyes that were only separated by 255 bp in the no-alternative-exon biopolymer shown at the top of the figure.

[0037] Figure 2B is a table showing the size of each of the exons (alternative exons are shown by shaded blocks) present in the MAPT gene. Figure 2A illustrates, generally, the various splicing permutations possible in the MAPT gene. As shown in that figure, exons 2, 3 and 10 are considered "alternative" exons, and may – or may not – be present in MAPT mRNA.

[0038] The user may also suitably label third, fourth, or even additional locations on the biopolymer with additional labels (with, *e.g.*, labeled nucleotides). Such additional labels may include the same fluorophore as the first or second labels, or may include fluorophores distinct from those on the first and second labels. The user may then correlate the distance between the third label and the first label, the third label and the second label, or both, to the presence or absence of an alternative exon disposed between the third label and the first label, between the third label and the second label, or both. The correlation may also provide the relative positions of one or more labels.

[0039] This is also shown by Figure 2. In embodiment labeled "2 + 10", the biopolymer includes alternative exons 2 and 10, which exons are disposed between the first and second and second and third labels (reading from left to right) on the biopolymer. The user can then determine the presence (or relative positions) of these exons by comparing the distances between the labels on the "2+10" embodiment against the distances between the labels on the "zero" embodiment shown at the top of the figure.

[0040] In addition to gleaning information about the structure of the biopolymer under study from the distance between the labels, the user can also obtain structural information based on the relative order of two or more probes, which is facilitated by the probes bearing differently-colored fluorophores. For example, if three probes (red, yellow, and green) are used, a sequence to which the probes bind in the order red – yellow – green is structurally different from a sequence to which the probes bind in the order yellow – red – green. Thus, the user may glean information about a sample by observing both the relative order in which probes are bound/arranged on the sample as well as the relative distances between probes.

[0041] Returning to the nonlimiting example described above, the user that compares two samples could determine – by accounting for the relative order of the probes and the distances between the probes – that two samples differ in (1) the order in which certain nucleotide sequences appeal (evidenced by probes being in different orders on the different samples) and (2) the number of, e.g., copy variations in a given sample (evidenced by certain probes being father apart on one sample than on another).

[0042] The labels are suitably separated from one another by about 30 bp to about 1000 bp, but more suitably about 30 bp. As described elsewhere here, a number of techniques (*e.g.*, SHRIMP, FIONA, SHREC, or other techniques known to those of ordinary skill in the art) enable resolution of labels separated from one another by small distances on the order of only hundreds or even tens of base pairs.

[0043] In another aspect, the present invention provides methods of obtaining structural information about a DNA sample. These methods suitably include nicking a first double-stranded DNA sample with a sequence-specific nicking endonucleoase. Such "nickases" are known the art, and are available from, *e.g.*, New England Biolabs (www.neb.com).

[0044] The methods suitably include incorporating one or more dye-labeled nucleotides at two or more nicking sites effected by the nicking endonuclease. Depending on the endonuclease and the sample under analysis, the nickage may effect one, two, or multiple nicking sites along the length of the sample. The labeled nucleotides are suitably incorporated into the biopolymer via a polymerase. The labeled nucleotides are, in suitable embodiments,

terminator nucleotides that counteract the effect of the polymerase and do not promote further chain lengthening. The nucleotides may bear the same fluorophore label or different labels, depending on the needs of the user.

[0045] The methods also suitably include linearizing a portion of the first double-stranded DNA sample that includes at least two dye-labeled nucleotides. Once the labeled DNA is linearized, the user may then register or otherwise account for the positions of two or more labeled dye-labeled nucleotides for use in further analysis.

[0046] One such analysis includes correlating the relative positions of two or more dye-labeled nucleotides to one or more structural characteristics of the first double-stranded DNA sample. This may entail – as shown in Figure 9 – determining the distance between two labels that are known to flank a region of interest, such as a region known to contain a certain mutation or copy number variation in some individuals. By comparing the between-labels distance on the sample to the between-labels distance on a control sample (or the between-labels distance on another sample taken from another individual or individuals), the user can determine whether the subject under analysis may have (or not have) a particular mutation.

[0047] In some embodiments, the "barcode" derived from the relative positions of the labels present on the biopolymer sample provides information regarding the relative position of the first double-stranded DNA sample within a principal double-stranded DNA sample from which the first double-stranded DNA sample was derived. The term "barcode" means a set of signals (e.g., from fluorescent labels spaced apart from one another) that represent a structural characteristic of a sample (e.g., the distance between two labels may be correlated to the presence of an extra copy of a gene in the region between the labels). The "barcode" can also be used to identify a particular sample where the set of signals from labels disposed on the sample is unique to that sample or distinguishes that sample from other samples under study.

[0048] For example, a user may determine that a portion of the barcode on a first sample taken from a "parent" sample overlaps with the barcode on a second sample taken from the "parent" sample, thus indicating that the "parent" sample included the region common to the first and second samples. Such "parent" samples may be digested to give rise to smaller oligonucleotides, which can then be themselves analyzed by the various methods described herein and then, by "barcoding" the smaller oligonucleotides, the user can then determine the relative positions of the oligonucleotides in the "parent" sample.

[0049] This is shown in Figure 13, which depicts (graphically) the steps of digesting a parent sample of DNA, placing barcodes on the products that result from the digestion, and the alignment – suitably done by computational methods – of products having corresponding

barcodes so as to piece together the parent and the effective barcode for the parent. In this way, the user can then correlate the barcode on the parent to, for example, physiological conditions in a subject. This can be done where the restriction enzymes used to digest the parent are known to isolate genomic regions that may contain copy number variations, exons, or other mutations that can be detected by comparing the distance between two labels disposed on the region of interest to the distance between two labels that are disposed on a "control" or "standard" that is known to lack (or to possess) the mutation or exon of interest.

[0050] As a non-limiting example, the user may place – by methods described here – a barcode of labels on the digestion products of a "parent sample" and then computationally reassemble those products to reform the "parent," with barcode. The user can then compare the barcode of the "parent" to other known samples to determine one or more characteristics of the parent, such as copy number variations, addition or deletion of exons, and the like. In this way, the user can perform a qualitative assessment of a "parent" sample by, effectively, placing all of the digestion products and their barcodes in their proper context within the "parent."

[0051] The methods can suitably include nicking a second double-stranded DNA sample with a sequence-specific nicking endonucleoase, incorporating one or more dye-labeled nucleotides at two or more nicking sites effected by the nicking endonuclease, linearizing a portion of the second double-stranded DNA sample that includes at least two dye-labeled nucleotides, and registering (e.g., recording or noting) the relative positions of two or more labeled dye-labeled nucleotides.

[0052] These relative positions – i.e., the barcode – of the labels can (as previously described) be used to determine the relationship between the first and second double-stranded DNA samples in a primary double-stranded DNA sample from which the first and second double-stranded DNA samples were derived.

[0053] In some embodiments, the user compares the relative positions of the two or more dye-labeled nucleotides to the positions of the same dye-labeled nucleotides on a second double-stranded DNA sample contacted with the same nicking endonucleoase. In this way, the user can compare the "barcodes" on different samples taken from different sources. This enables a qualitative comparison between multiple samples, as shown in Figure 10. In that figure, samples from Subjects A, B, and C and processed according to the claimed methods. As shown, Subject C's sample lacks a label that bound to the samples from Subjects A and B, suggesting that Subject C's DNA lacks that particular region. The user may then correlate this deleted region to a physiological characteristic of Subject C, or may compare Subject C's results to the

results of still other subjects to identify those characteristics common to individuals missing that region of DNA.

[0054] Also provided are methods of obtaining sequence information about a nucleic acid biopolymer. These methods suitably include binding a first fluorescently labeled sequence specific probe having a first binding sequence to a single-stranded nucleic acid biopolymer. This is shown in, e.g., Figure 11. The user then contacts the single-stranded nucleic acid biopolymer with a first terminator nucleotide bearing a fluorescent label A (e.g., adenine bearing Cy5), with a second terminator nucleotide bearing a fluorescent label B (e.g., cytosine bearing Alexa 405), with a third terminator nucleotide bearing a fluorescent label C, and with a fourth terminator nucleotide bearing a fluorescent label D. The user then illuminates the nucleic acid biopolymer so as to determine the presence (or relative positions) of the first terminator nucleotide, the second terminator nucleotide, the third terminator nucleotide, the fourth terminator nucleotide, or any combination thereof, adjacent to the first labeled sequence-specific probe.

[0055] The binding sequence of the first probe is suitably between 4 and 6 nucleotides. In some embodiments, the fluorescent labels of the nucleotides have different excitation wavelengths. In others, two or more of the labels share an excitation wavelength. The excitation nwavenength of a labeled nucleotide may be the same – or different – from the excitation wavelength of the labeled, sequence-specific probe.

[0056] The methods also suitably include contacting at least four fluorescently labeled probes having, respectively, second, third, fourth, and fifth binding sequences to the single-stranded nucleic acid biopolymer. The second binding sequence is suitably constructed by eliminating the base at the 5' end of the first binding sequence and adding a first replacement base to the 3' end of the first binding sequence.

[0057] Similarly, the third binding sequence is constructed by eliminating the base at the 5' end of the first binding sequence and adding a second replacement base to the 3' end of the first binding sequence. The fourth binding sequence is suitably constructed by eliminating the base at the 5' end of the first binding sequence and adding a third replacement base to the 3' end of the first binding sequence, and the fifth binding sequence is constructed by eliminating the base at the 5' end of the first binding sequence and adding a fourth replacement base to the 3' end of the first binding sequence. These probes suitably bear different fluorophores from one another, and may bear different fluorophores than the first probe.

[0058] As a non-limiting example, the first probe may comprise the sequence 5'-CTAGC-3'. In the second cycle of probing, the C at the 5' end of the probe is eliminated, and the T then becomes the 5' end of the probe, with the 3' end of the probe being as follows: 5'

TAGCA-3'; 5'-TAGCT-3'; 5'-TAGCG-3'; 5'-TAGCC-3'. These labeled probes are then contacted to the biopolymer, and by illuminating the probes with the appropriate excitation wavelength, the user may determine the location of the new probes and thus obtains information regarding the sequence of the biopolymer under study. While the binding sequence shown in this example is 5 bp in length, binding sequences are suitably from 1 to 100 bp in length, but more suitably from 4 bp to 6 bp in length.

[0059] The methods also suitably include illuminating the nucleic acid biopolymer so as to determine the presence (or relative positions) of the first terminator nucleotide, the second terminator nucleotide, the third terminator nucleotide, the fourth terminator nucleotide, or any combination thereof, adjacent to the second labeled sequence-specific probe.

[0060] Figure 11 is one non-limiting embodiment of the methods. As shown in that figure, the user may bind first and second probes – having different binding sequences – to the biopolymer sample. The user then contacts the sample with labeled nucleotides under such conditions that only a single nucleotide binds to the single-stranded DNA, adjacent to the bound probe. This gives rise to a given probe-nucleotide pair displaying two labels, which labels may – as shown in the figure – be different from one another. The user can then illuminate the sample as needed to visualize or otherwise locate the probe-nucleotide pairs. Probes and nucleotides may be joined by ligases. In some embodiments, there may be a gap (1+ bps) between the probe and the nucleotide, which gap can be filled by a polymerase and a supply of nucleotides, which nucleotides may themselves be labeled. Ligase may also be used to join to probes, with the gap being 'filled in' by labeled nucleotides. Non-fluorescent probes may be used.

[0061] The user may, after completing a first cycle of probe-binding followed by binding of labeled nucleotides, begin a second cycle using probes that consider the sequence information learned in the first cycle. For example, a first probe may have a sequence of AAGG, and the labeled nucleotide that binds adjacent to the probe is T. In the next cycle, the user may take advantage of this information and use a probe that has a sequence of AGGT, so as to obtain additional sequence information, as described above.

[0062] In another aspect, the present invention provides methods of obtaining structural information about a nucleic acid biopolymer. These methods suitably include (a) contacting a double-stranded biopolymer with a nicking endonuclease so as to give rise to at least two nicking sites; (b) contacting the at least two nicking sites with a first nucleotide bearing a fluorescent label A (e.g., Cy3); (c) removing excess first nucleotide; (d) illuminating the double-stranded biopolymer so as to determine the presence or relative positions of the first nucleotide; (e) contacting the at least two nicking sites with a second nucleotide bearing a fluorescent label B

(e.g., Cy5) and (f) removing excess second nucleotide. The user suitably illuminates double-stranded biopolymer so as to determine the presence or relative positions of the second nucleotide.

[0063] The user suitably contacts at least two nicking sites with a third nucleotide bearing a fluorescent label C (e.g., Alexa 405); removes excess third nucleotide; (j) illuminates the double-stranded biopolymer so as to determine the presence or relative positions of the third nucleotide. The methods also include (k) contacting the at least two nicking sites with a fourth nucleotide bearing a fluorescent label D, (*l*) removing excess fourth nucleotide; and (m) illuminating the double-stranded biopolymer so as to determine the presence or relative positions of the first nucleotide.

[0064] In this way, the nickase "opens" the double-stranded sample so as to make available a nucleotide adjacent to the location where the nickase binds. The user then introduces the first labeled nucleotide (e.g., cytosine), and assays the biopolymer to determine whether and where that nucleotide may have bound. This is then repeated with the other nucleotides (guanine, tyrosine, adenosine), following the introduction of each of which the user assays (via illumination) for binding of each newly-introduced nucleotide.

[0065] The preceding steps (identified as (b) through (m)) may then be repeated so as to enable the user to obtain additional sequence information with the addition of each successive labeled nucleotide.

[0066] The illumination also suitably establishes the relative positions of one or more of the labeled nucleotides. At least a portion of the sample bearing two or more labels is suitably linearized for this analysis. The user then determines the distances between the two or more labeled nucleotides residing within the linearized portion of the double-stranded biopolymer. These distances may then be used to arrive at a barcode for the sample under analysis.

[0067] In some variations, the user may induce a second nicking site adjacent to the terminator nucleotide residing at the first nicking site. The user suitably contacts the second nicking site with a first nucleotide bearing a fluorescent label A, with a second nucleotide bearing a fluorescent label B, with a third nucleotide bearing a fluorescent label C, and with a fourth nucleotide bearing a fluorescent label D, and illuminating the doublestranded biopolymer so as to determine the labeled nucleotide incorporated at the second nicking site.

[0068] This is shown in Figure 12. As shown in that figure, two nickase molecules bind to a double-stranded DNA sample and effect a nicking site at their ends, shown by the boxed "N" in the figure. The user then introduces labeled nucleotides in sequence. As shown in the figure, adenosine is introduced first and binds to the T located on the DNA strand opposite

the left-hand probe. Because there is an adenosine opposite to the right hand probe, the labeled adnensine does not bind at that site, and an "X" signifies that there was no binding upon introduction of the first labeled base. Additional nickases and labeled bases are introduced, and the user is able to sequence the biopolymer target by sequential addition of labeled bases following by illumination of the labeled sample. The sequence information gleaned from the method can then be used to design probes that bind to particular sequences, which probes can then be used to "barcode" a given sample for further characterization., such as comparing the relative distances between two or more labeled probes on a first sample to the distances to the corresponding labeled probes on a different – or control – sample.

[0069] The invention also provides kits for performing multiplexed hybridization. These kits suitably first include a plurality of hybridization probes. Each of the probes is suitably of a different color or responds to a different excitation wavelength. The kits also suitably includes instructions for applying at least two of these hybridization probes to a nucleic acid sample, for linearizing the labeled sample, and for imaging at least one of the hybridized nucleic acids. In some embodiments, the user images two or more hybridized probes so as to determine the distance between the two probes or the relative positions of the two probes.

[0070] Depending on certain conditions, the user may populate the enture biopolymer region between adjacent nicking sites with labeled nucleotides. This is suitably accomplished when the nicking sites are comparatively close to one another. Under illumination, biopolymer regions that bear at least some labeled nucleotides are comparatively bright; regions that lack labeled nucleotides are comparatively dark. The user, however, may nonetheless glean information from both bright and dark regions.

[0071] So-called bright regions provide sequence information, as the user can illuminate the region with the excitation wavelengths that correspond to the various labeled nucleotides disposed within the region. In other embodiments, the user can, by determining the distance between bright regions (or even nucleotides) that flank a dark region, assess whether the dark region – by virtue of its size -- comprises copy number variations, exons, or other structural features of interest. Thus, structural information can be gleaned from both bright and dark regions.

[0072] In some embodiments, the user may elect to utilize nickases that have binding sequences complementary to a region on the biopolymer sample that is of particular interest. In this way, the user can efficiently obtain sequence information for only that region (or regions) believed to be of greatest interest or importance.

[0073] The user may also suitably determine the sequence of at least a portion of the biopolymer sample by correlating the order of fluorophores visible under illumination to the nucleotides to which one or more of the fluorophores correspond.

[0074] **Additional Dislosure**

[0075] **Imaging techniques**

[0076] Several techniques improve optical resolution in fluorescent imaging by at least one order of magnitude. Application of these imaging techniques to single molecule DNA and RNA analyses vastly accelerates the applications discussed above.

[0077] One such technique, termed Fluorescence Imaging with One Nanometer Accuracy (FIONA), involves the localization of single organic fluorophores by fitting a distribution function to the light collected from the fluorophore. The center of this distribution can be localized with 1.5 nm precision. FIONA has been used to study the translocation of molecular motors or to measure small distances.

[0078] Extensions of this technique include Single molecule-High Resolution Imaging with Photobleaching (SHRIMP) which is able to resolve adjacent fluorophores of the same color with about 10 nm resolution. FIONA has been extended to two colors, developing a method termed single-molecule high-resolution colocalization (SHREC). Users might, for example, colocalize Cy3 and Cy5 dyes as close together as 10nm, which dyes can be attached at the ends of a short DNA. Also useful is a method of multicolor stochastic optical reconstruction microscopy (STORM), which allows combinatorial pairing of reporters and activators. Iterative, color-specific activation of sparse subsets of these probes allows localization with nanometer accuracy.

[0079] **Genome mapping methods**

[0080] Structural variations play a very important role in human health and common diseases. These variations are defined as being longer than 1kb. But despite their importance, most genome-wide approaches for detecting copy number variations (CNVs) are indirect, depending on signal intensity differences between samples and controls to predict regions of variation. Such approaches therefore provide limited quantitative signal and positional information, and cannot detect balanced events such as inversions and translocations. For example, microarray-based platforms including SNP array, oligo Comparative Genomic Hybridization (CGH) array, and BAC CGH arrays are the main techniques for structural variations discovery. Non-uniform sensitivity, specificity, and probe density of these platforms often lead to conflicting results even with identical samples. This qualitative measurement requires further confirmation by low throughput detection methods, such as PCR and FISH.

[0081] Optical mapping

[0082] The single molecule techniques described above are well suited for studying structural variations. However, due to the optical nature of the mapping, they are limited in their ability to resolve motifs that are closer than about ~1 kbp. Significantly greater mapping efficiency can be achieved by resolving features less than 100 bp apart. In turn, this substantially improves our ability to identify structural variations in native, long genomic DNA molecules.

[0083] A suitable mapping scheme is based on the labeling of sites generated by nicking endonucleases. A nicking endonuclease with a five base recognition sequence will, on average, generate a 1 kb physical map across the whole genome. Based on *in silico* whole genome mapping, a large portion of such nicking sites fall within 1000 bp of each other, which distance which cannot be resolved with conventional optics. This reduces map resolution and makes map assembly more difficult.

[0084] An example is the recognition sequences (motifs) for two commercially available nicking endonucleases ranging with 5 base to 7 base recognition sites. An algorithm to map all the nicking sites against the human reference genome was designed.

[0085] In the case of enzyme Nt.BstNB1 (5 base motif GACTC), there are 2.1×10^6 sites across whole human genome, which produces an average of 1.5 kb between nicks. For enzyme Nt.BspQ1 (7 base motif GCTCTTC), there are 2.2×10^5 nicking sites separated on average by 15 kbp. In principle, the nicking sites using the 5 base motif are resolvable with conventional optics (~1 kbp), but in silico analysis revealed that almost half the nicking sites fall within 1kbp of each other, rendering them indistinguishable from one another. Using the 7 base motif, one can resolve a greater number of sites. As discussed below, this leads to challenges in uniquely mapping a fragment of DNA.

[0086] Improved resolution in DNA mapping

[0087] *In silico* mapping was used to determine the percentage of DNA fragments that can be uniquely mapped based on currently available nicking enzymes and our existing optical detection system.

[0088] Figure 1A shows the results for the nicking endonuclease Nt.BstNBI (5 base motif). For 1000 bp optical resolution, only about 12% of fragments can be uniquely identified with 8 nicking sites. On the other hand, to achieve 100 bp resolution, over 97% of the fragments are unique. Closely clustered nicking sites pack more sequence information and their distributions are unique. Furthermore, with only 8 nicking sites, one merely needs a 12 kb fragment (on average) to enable unique mapping of the fragment to the reference genome.

[0089] The nicking map for enzyme Nt.BspQI (7 base motif) (Fig. 1B) shows that by improving the resolution to 100 bp, one gains very little because fewer Nt.BspQI nicking sites fall within 1kbp of each other. On average 8 consecutive Nt.BstQI nicking sites are needed to uniquely identify a DNA fragment using this enzyme but the average size of fragments is about 120 kb. There are significant regions of the genome (~30%) that cannot be mapped due to the lack of consecutive nicking sites within a length of DNA that can be reasonably extracted with existing methods.

[0090] Without being bound to any single theory, some advantages of the claimed invention can be identified. First, much more information about a DNA fragment is available when resolving closely spaced nicking sites. The ability to uniquely map a fragment to the genome is vastly improved.

[0091] Second, with improved resolution, one may resolve much smaller structural variations than is currently possible with optical methods. Finally, improved resolution also helps us identify large scale structural variations.

[0092] **Additional Background on the Figures**

[0093] In Fig. 1C is shown an example of a fragment having a 150 kbp insertion. Successfully mapping the fragment (and thus identifying the location of the insertion within the genome), can use a contiguous set of 8 nicking sites adjacent to the insertion. With limited optical resolution, this necessitates large (>300 kbp) genomic fragments. These are difficult to generate with standard DNA extraction protocols. In contrast, with 100 bp resolution, one may employ a dense nicking site distribution using a fragment only slightly larger than the insertion to uniquely map the fragment.

[0094] **The need for high throughput digital profiling of alternative transcriptome**

[0095] Another nucleic acid analysis that can greatly benefit from improved mapping capability is alternative splicing of RNA. During pre-RNA splicing, introns are removed, and exons are joined together to form mature RNA. Alternative splicing is the process by which a single primary transcript yields different mature RNAs. This leads to the production of protein isoforms with diverse and even antagonistic functions. Recent studies showed the large proteomic complexity and diversity are achieved with a limited number of genes. In human genome, ~75% of human genes exhibit alternative splicing. While the human genome contains 25,000 genes, it can produce several hundred thousand different types of proteins through alternative splicing.

[0096] Alternative splicing variants of many genes have a critical impact on all major aspects of cell biology, including cell cycle control, apoptosis and more. Aberrant splicing has

been found to be associated with various diseases, including cancer, and recent studies suggest that mRNAs are more frequently alternatively spliced in cancerous tissues than in normal ones. Other examples include the significant reduction of the full-length transmembrane conductance regulator (CFTR) gene due to aberrant exon inclusion and inclusion which gives rise to atypical forms of Cystic Fibrosis. Another example is the microtubule associated protein Tau (MAPT gene). MAPT is required for the polymerization and stability of microtubules as well as axonal transport in neurons. Aberrant splicing of Tau exon 10 leads to the development of neurodegenerative disease, dementia FTDP-17.

[0097] A number of techniques have been developed to quantify RNA splicing variants. First, oligo microarray and fiber-optic arrays have been used for globally detecting gene splicing variants. However, because small fragments of full RNA transcripts are interrogated one at a time in array technology, only one splicing event (two exons at a time) can be detected at a time. Thus, it is difficult to quantify how many exons are included or excluded in one specific splicing variant. Furthermore, non-specific hybridization can result in many false positives which require further confirmation.

[0098] Second, real-time PCR can obtain splicing information by quantifying one exon junction at a time but is limited by stringent reaction conditions, low throughput, and high cost. Third, so-called next generation sequencing technologies have been employed in digital gene expression profiling and could be used in profiling alternative splicing variants. However, they are largely based on short sequence reads and have the same limitations as microarrays with regards to full-length RNA samples.

[0099] A disadvantage common to existing transcriptome-focused technologies is that none is capable of monitoring combinations of alternatively spliced exons, as they occur within individual transcripts. Under existing methods, exon exclusion is hard to confirm, which may result in false exclusion of certain exons.

[0100] Despite the enormous importance of alternative splicing to mammalian biology, current solutions to deciphering this problem face challenges. Indeed, little is known about how alternative splicing is regulated and coordinated through the developmental stage due to a lack of robust methods to quantify RNA splice variants.

[0101] Improving resolution beyond conventional optical limitations

[0102] As an example of the advantages available to improved resolution, one can consider an optical barcoding approach for microtubule associated protein Tau (MAPT) gene which is required for the polymerization and stability of microtubules as well as axonal transport

in neurons. Aberrant splicing of Tau exon 10 leads to the development of neurodegenerative disease such as dementia FTDP-17.

[0103] An exemplary RNA barcoding scheme is shown in Fig. 2. Three exons (2, 3, and 10) in MAPT transcripts can undergo alternative splicing, exon 2 and exon 3 are always spliced together. Thus, six different MAPT transcripts can be generated by alternative splicing. The MAPT gene structure is shown in Fig. 2A.

[0104] All six possible alternative splicing isoforms are indicated (Zero, 2, 10 2+10, etc.), and the length of each exon is indicated in Fig. 2B. Conventional optical resolution unable to discriminate labels associated with different exons. If the position of the exons could be resolved, the measured distance between the labels will identify each splicing variant in a manner similar to reading a barcode.

[0105] To form a barcode in this example, four exon specific oligo probes could be designed to specifically hybridize to exon 1 (Cy3-green), exon 7(Cy5-red), exon 11(Cy5-red), exon 13 (Cy3-green) respectively, as shown with green and red arrows in Fig. 2C. The distance between the labels can be used to identify which variant is present and the color sequence (i.e. Green-Red-Red-Green) indicates the presence of a fully labeled transcript. Further, the disclosed barcoding scheme is easily multiplexed.

[0106] For example, if the same two colors (*e.g.*, green and red) with four different probes were used to tag a different gene, a color sequence can be designed for this particular gene that is different from that of the MAPT gene. The sequence of color can thus be used to define the specific gene and the distance between the labels of that color sequence determine the individual splicing variant of that specific gene. In this two-color, four-probe approach, there are $2^4 = 16$ different color sequences to interrogate 16 different genes simultaneously with unlimited power for splicing variants. If 4 colors of 8 different probes were used, 4^6=65536 different genes can be investigated simultaneously, which is more than the entire human transcriptome (figure 2C).

[0107] This approach has three important advantages over current expression profiling technologies for interrogating RNA splicing: (i) By mapping the distribution of exons within a single transcript concurrently, one can determine the relationships amongst multiple alternatively spliced exons within the same transcript. (ii) The digital nature of the barcoding scheme means not only can one quantify the individual splicing variant, one can quantify the total gene expression by adding up the all the splicing variants. (iii) The barcoding scheme will provide maximum multiplex detection capability. Realizing these advantanges necessitates an imaging technology with resolution that far exceeds conventional optical approaches.

[0108] The need for low cost and high throughput whole genome sequencing

[0109] The success of the Human Genome Project (HGP) is largely due to the continuous development of Sanger sequencing method through parallelization, automation, miniaturization, better chemistry and informatics. As the workhorse of the Human Genome Project, Sanger sequencing method has dominated the DNA sequencing field for nearly three decades, and its 800 Q20 base read length is significant.

[0110] These newly emerging sequencing technologies can be grouped into two categories based on the detection methods, sequencing either by ensemble detection or by single molecule detection. Since multiple DNA copies are needed in ensemble detection, the genetic information, such as haplotype and RNA splicing pattern is lost during the process. While sequencing by single molecule detection may be able to recover haplotype information, the read length of current single molecule sequencing method (e.g., Helicos tSMS) is 50 bp or less, which is far shorter than the average distance of 1 kbp between two SNPs. Thus, as with the predecessor Sanger sequencing method, critical genetic information such as haplotypes and RNA splicing pattern is still difficult to obtain with these "next generation" sequencing technologies. The present invention, among other things, effects DNA sequencing length over 10kb.

[0111] Sequencing by hybridization is a well known method that employs microarray-based hybridization assays to determine the sequence of nucleic acid molecules. Normally, short oligos with known sequence (<100mer) constructed on a microarray are used to capture (i.e., hybridize) and interrogate the target molecules. The microarray assays produce a list of all subsequences of hybridized oligos found at least once in the target molecules. However, the list does not reveal the locations of the sequences of hybridized oligos or nor does the list provide the number of the times an oligo may be present on a target molecule. The present invention, however, obtains such information.

[0112] Figure 3 displays the starting materials for sequencing. A set of 5-mer (i.e., five nucleotides in length) oligos with 5' end labeled with different color fluorophores; 4 nucleotide terminators labeled with different color fluorophores; arrays of linearized single stranded DNA molecules, or double stranded DNA molecules with partial ssDNA gaps.

[0113] Figure 4 describes the first cycle of an exemplary sequencing reaction. After the first cycle, each hybridization and incorporation events are recorded and localized alone linearized DNA molecules by STORM imaging technique. The probes are then washed away. In the next cycle, 4 more 5-mer probes AGTCA, AGTCT, AGTCG, and AGTCT are introduced and hybridize on the same locations as previous probes, as they share the same sequences as previous probes. A polymerase then incorporates the nucleotide terminators (Figure 5).

[0114] This process is adapted to be multiplexed (using labels of different colors) and to produce large number of sequences read during one cycle (Figure 6). Also developed are algorithms to prioritize the sequential addition of 5-mer probes. The super imaging techniques used here included SHRIMP, SHREC, STORM.

EXAMPLES

[0115] Single-molecule high resolution co-localization (SHREC) and single-molecule high-resolution imaging with photobleaching (SHRImP) methods have been developed to measure distances between two fluorophores that are closer than Rayleigh limit (≈250 nm for visible excitation).

[0116] Combining the two techniques adds another dimension to the power of localization methodology and tens of distances could potentially be resolved by using several fluorophores of different colors each having multiple members. To apply this to DNA, double-stranded DNA was stretched on a Polyacrylic acid and Polyallylamine coated surface, making the DNA relatively straight. To test SHRIMP, a DNA construct was made with a biotin followed by three Cy-3's at positions 475 bp, 172 bp, and 94 bp, corresponding to distances between Cy3 of 32nm, 58nm, and 90nm (figure 7B).

[0117] Additional detail is provided in Figure 7A. One PCR primer was labeled at 5' end with cy3 and the other primer was phosphorelated at 5' end. After PCR reaction, the 5' end of cy3 protect that strand from digestion by lambda exonuclease, which resulting in a single stranded DNA molecules. Once the single-stranded DNA molecules were generated, primer extension reactions were performed to introduce fluorescent dyes at each specific sequence positions. In this case, two short oligos with cy3 at their 5' end were hybridized respectively at 94 bp and 256 bp from one end. Another short oligo with a biotin at its 5' end was hybridized at the 3' end of the single stranded template. After extension by polymerase, the single stranded template was converted to double stranded DNA molecules and two cy3 dye molecules were introduced at specific locations.

[0118] Distances of 27 nm, 61nm, and 95 nm were measured, in excellent agreement with the expected distances. To test simultaneously SHRIMP and SHREC, Cy5 was placed at position zero, and two Cy3's at position 94 bp and position 172 bp, with their positions measured using a dual-view imaging system. The distances between Cy3-Cy5 pairs were 37±5 nm (32 nm expected) and 91±5 nm (87 nm expected), and the distance between Cy3-Cy3 pair to be 56±3 nm (58 nm expected) (figure 8). The agreement was excellent.

Claims

1. A method of assaying for the relative position of one or more exons, comprising:

labeling a biopolymer sample at a plurality of locations with a plurality of labels, wherein said labels are separated from one another by 30 base pairs to 1000 pairs on said biopolymer;

said plurality of labels including at least a first and second label located, respectively, at first and second locations on a biopolymer sample, such that the first and second labels flank a first region of the biopolymer sample that includes at least one constant exon;

determining the distance between the first and second labels; and

correlating the distance between the first and second labels to the presence, absence or relative position of an alternative exon in said first region of the biopolymer.

2. The method of claim 1, wherein correlating the distance between the first and second labels to the presence, absence, or relative position of said one or more alternative exons comprises comparing the distance between the first and second labels present on the biopolymer sample to the distance between labels that flank the first region of the biopolymer known not to contain the alternative exon.

3. The method of claim 1, wherein the correlating comprises comparing the distance between the first and second labels to the distance between the first and second locations on a biopolymer lacking the alternative exon between the first and second locations.

4. The method of claim 1, further comprising linearizing at least a portion of the biopolymer sample wherein the linearized portion of the biopolymer sample comprises at least two labels.

5. The method of claim 1, further comprising correlating the relative order of two or more probes to one or more structural characteristics of the biopolymer sample.

6. The method of claim 1, wherein the biopolymer comprises DNA complementary to an mRNA.

7. The method of claim 1, wherein the first and second labels comprise the same fluorophore.

8. The method of claim 1, wherein the first and second labels comprise different fluorophores.

9. The method of claim 1, wherein the plurality of labels includes at least a third label located at a third location on the biopolymer sample.

10. The method of claim 9, wherein the third label comprises the same fluorophore as the first label, the second label, or both.

11. The method of claim 9, wherein each of the first, second and third labels comprises a different fluorophore.

12. The method of claim 9, further comprising correlating the distance between the first and the third labels, the distance between the second and the third labels, or both, to the presence, absence, or relative position of an alternative exon disposed between the first and the third labels, an alternative exon disposed between the second and the third labels, or both.

13. The method of claim 1, wherein labeling the biopolymer sample comprises contacting the biopolymer sample with a nicking endonuclease.

14. The method of claim 13, wherein labeling the biopolymer sample comprises incorporating one or more dye-labeled nucleotides at two or more nicking sites effected by the nicking endonuclease.