COMPOSITIONS AND METHODS FOR ACCURATELY IDENTIFYING MUTATIONS

CROSS REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of U.S. provisional patent application Ser. No. 61/600,535, filed Feb. 17, 2012, which is incorporated herein by reference in its entirety.

STATEMENT REGARDING SEQUENCE LISTING

[0002] 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 360056_409 WO_SEQUENCE_LISTING_.txt. The text file is 4 KB, was created on February 14, 2013, and is being submitted electronically via EFS-Web.

BACKGROUND

1. Technical Field

[0003] The present disclosure relates to compositions and methods for accurately detecting mutations using sequencing and, more particularly, uniquely tagging double stranded nucleic acid molecules such that sequence data obtained for a sense strand can be linked to sequence data obtained from the anti-sense strand when obtained via massively parallel sequencing methods.

2. Description of Related Art

[0004] Detection of spontaneous mutations (e.g., substitutions, insertions, deletions, duplications), or even induced mutations, that occur randomly throughout a genome can be challenging because these mutational events are rare and may exist in one or only a few copies of DNA. The most direct way to detect mutations is by sequencing, but the available sequencing methods are not sensitive enough to detect rare mutations. For example, mutations that arise de novo in mitochondrial DNA (mtDNA) will generally only be present in a single copy of mtDNA, which means these mutations are not easily found since a mutation must be present in as much as 10-25% of a population of molecules to be detected by sequencing (Jones et al., *Proc. Nat'l. Acad. Sci. U.S.A.*105:4283-88, 2008). As another example, the spontaneous somatic mutation frequency in genomic DNA has been estimated to be as low as 1×10^{-8} and 2.1×10^{-6} in human normal and cancerous tissues, respectively (Bielas et al., *Proc. Nat'l Acad. Sci. U.S.A.* 103:18238-42, 2008).

[0005] One improvement in sequencing has been to take individual DNA molecules and amplify the number of each molecule by, for example, polymerase chain reaction (PCR) and digital PCR. Indeed, massively parallel sequencing represents a particularly powerful form of digital PCR because multiple millions of template DNA molecules can be analyzed one by one. However, the amplification of single DNA molecules prior to or during sequencing by PCR and/or bridge amplification suffers from the inherent error rate of polymerases employed for amplification, and spurious mutations generated during amplification may be misidentified as spontaneous mutations from the original (endogenous unamplified) nucleic acid. Similarly, DNA templates damaged during preparation (ex vivo) may be amplified and incorrectly scored as mutations by massively parallel sequencing techniques. Again, using mtDNA as an example, experimentally determined mutation frequencies are strongly dependent on the accuracy of the particular assay being used (Kraytsberg et al., *Methods* 46:269-73, 2008)—these discrepancies suggest that the spontaneous mutation frequency of mtDNA is either below, or very close to, the detection limit of these technologies. Massively parallel sequencing cannot generally be used to detect rare variants because of the high error rate associated with the sequencing process—one process using bridge amplification and sequencing by synthesis has shown an error rate that varies from about 0.06% to 1%, which depends on various factors including read length, base-calling algorithms, and the type of variants detected (see Kinde et al., *Proc. Nat'l. Acad. Sci. U.S.A.* 108:9530-5, 2011).

BRIEF SUMMARY

[0006] In one aspect, the present disclosure provides a double-stranded nucleic acid molecule library that includes a plurality of target nucleic acid molecules and a plurality of random cyphers, wherein the nucleic acid library comprises molecules having a formula of X^a-X^b-Y, X^b-X^a-Y, Y-X^a-X^b, Y-X^b-X^a, X^a-Y-X^b, or X^b-Y-X^a (in 5' to 3' order), wherein (a) X^a comprises a first random cypher, (b) Y comprises a target nucleic acid molecule, and (c) X^b comprises a second random cypher. Furthermore, each of the plurality of random cyphers comprise a length ranging from about 5 nucleotides to about 50 nucleotides (or about 5 nucleotides to about 10 nucleotides, or a length of about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, or about 20 nucleotides).

[0007] In certain embodiments, the double-stranded sequences of the X^a and X^b cyphers are the same (e.g., X^a=X^b) for one or more target nucleic acid molecules, provided that each such target nucleic acid molecule does not have the same double-stranded cypher sequence as any other such target nucleic acid molecule. In certain other embodiments, the double-stranded sequence of the X^a cypher for each target nucleic acid molecule is different from the double-stranded sequence of the X^b cypher. In further embodiments, the double-stranded nucleic acid library is contained in a self-replicating vector, such as a plasmid, cosmid, YAC, or viral vector.

[0008] In a further aspect, the present disclosure provides a method for obtaining a nucleic acid sequence or accurately detecting a true mutation in a nucleic acid molecule by amplifying each strand of the aforementioned double-stranded nucleic acid library wherein a plurality of target nucleic acid molecules and plurality of random cyphers are amplified, and sequencing each strand of the plurality of target nucleic acid molecules and plurality of random cyphers. In certain embodiments, the sequencing is performed using massively parallel sequencing methods. In certain embodiments, the sequence of one strand of a target nucleic acid molecule associated with the first random cypher aligned with the sequence of the complementary strand associated with the second random cypher results in a measureable sequencing error rate ranging from about 10^{-6} to about 10^{-8}.