

637449

# THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

*June 30, 2014*

**THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.**

**APPLICATION NUMBER: 61/613,413**

**FILING DATE: *March 20, 2012***

**RELATED PCT APPLICATION NUMBER: *PCT/US13/32665***

**THE COUNTRY CODE AND NUMBER OF YOUR PRIORITY APPLICATION, TO BE USED FOR FILING ABROAD UNDER THE PARIS CONVENTION, IS *US61/613,413***



Certified by

*David J. Kyros*

Under Secretary of Commerce  
for Intellectual Property  
and Director of the United States  
Patent and Trademark Office

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

### Provisional Application for Patent Cover Sheet

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c)

#### Inventor(s)

Inventor 1

Remove

Given Name	Middle Name	Family Name	City	State	Country ;
Michael		Schmitt	Seattle	WA	US

Inventor 2

Remove

Given Name	Middle Name	Family Name	City	State	Country ;
Jesse		Salk	Seattle	WA	US

All Inventors Must Be Listed – Additional Inventor Information blocks may be generated within this form by selecting the **Add** button.

Add

#### Title of Invention

METHODS OF LOWERING THE ERROR RATE OF MASSIVELY PARALLEL DNA SEQUENCING USING DUPLEX CONSENSUS SEQUENCING

Attorney Docket Number (if applicable)

72227.8043.US00

#### Correspondence Address

Direct all correspondence to (select one):

The address corresponding to Customer Number

Firm or Individual Name

Customer Number

34055

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

No.

Yes, the name of the U.S. Government agency and the Government contract number are:

NIH RO1 CA115802; NIH RO1 CA102029

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

**Entity Status**

Applicant claims small entity status under 37 CFR 1.27

- Yes, applicant qualifies for small entity status under 37 CFR 1.27  
 No

**Warning**

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

**Signature**

Please see 37 CFR 1.4(d) for the form of the signature.

Signature	/Lara J. Dueppen/			Date (YYYY-MM-DD)	Mar 20, 2011
First Name	Lara J.	Last Name	Dueppen	Registration Number (If appropriate)	65002

This collection of information is required by 37 CFR 1.51. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 8 hours to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. **DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. This form can only be used when in conjunction with EFS-Web. If this form is mailed to the USPTO, it may cause delays in handling the provisional application.**

## Privacy Act Statement

**The Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that : (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

**METHODS OF LOWERING THE ERROR RATE OF MASSIVELY PARALLEL DNA SEQUENCING USING DUPLEX CONSENSUS SEQUENCING**

**STATEMENT OF GOVERNMENT INTEREST**

**[0001]** The present invention was made with government support under Grant Nos. RO1 CA115802 and RO1 CA102029 awarded by the National Institutes of Health. The Government has certain rights in the invention.

**BACKGROUND**

**[0002]** The advent of massively parallel DNA sequencing has ushered in a new era of genomic exploration by making simultaneous genotyping of hundreds of billions of base-pairs possible at small fraction of the time and cost of traditional Sanger methods [1]. Because these technologies digitally tabulate the sequence of many individual DNA fragments, unlike conventional techniques which simply report the average genotype of an aggregate collection of molecules, they offer the unique ability to detect minor variants within heterogeneous mixtures [2].

**[0003]** This concept of “deep sequencing” has been implemented in a variety of fields including metagenomics [3, 4], paleogenomics [5], forensics [6], and human genetics [7, 8] to disentangle subpopulations in complex biological samples. Clinical applications, such prenatal screening for fetal aneuploidy [9, 10], early detection of cancer [11] and monitoring its response to therapy [12, 13] with nucleic acid-based serum biomarkers, are rapidly being developed. Exceptional diversity within microbial [14, 15] viral [16-18] and tumor cell populations [19, 20] has been characterized through next-generation sequencing, and many low-frequency, drug-resistant variants of

therapeutic importance have been so identified [12, 21, 22]. Previously unappreciated intra-organismal mosaicism in both the nuclear [23] and mitochondrial [24, 25] genome has been revealed by these technologies, and such somatic heterogeneity, along with that arising within the adaptive immune system [13], may be an important factor in phenotypic variability of disease.

**[0004]** Deep sequencing, however, has limitations. Although, in theory, DNA subpopulations of any size should be detectable when deep sequencing a sufficient number of molecules, a practical limit of detection is imposed by errors introduced during sample preparation and sequencing. PCR amplification of heterogeneous mixtures can result in population skewing due to stochastic and non-stochastic amplification biases and lead to over- or under-representation of particular variants [26]. Polymerase mistakes during pre-amplification generate point mutations resulting from base mis-incorporations and rearrangements due to template switching [26, 27]. Combined with the additional errors that arise during cluster amplification, cycle sequencing and image analysis, approximately 1% of bases are incorrectly identified, depending on the specific platform and sequence context [2, 28]. This background level of artifactual heterogeneity establishes a limit below which the presence of true rare variants is obscured [29].

**[0005]** A variety of improvements at the level of biochemistry [30-32] and data processing [19, 21, 28, 32, 33] have been developed to improve sequencing accuracy. The ability to resolve subpopulations below 0.1%, however, has remained elusive. Although several groups have attempted to increase sensitivity of sequencing, several limitations remain. For example techniques whereby DNA fragments to be sequenced

are each uniquely tagged [34, 35] prior to amplification [36-41] have been reported. Because all amplicons derived from a particular starting molecule will bear its specific tag, any variation in the sequence or copy number of identically tagged sequencing reads can be discounted as technical error. This approach has been used to improve counting accuracy of DNA [38, 39, 41] and RNA templates [37, 38, 40] and to correct base errors arising during PCR or sequencing [36, 37, 39]. Kinde et. al. reported a reduction in error frequency of approximately 20-fold with a tagging method that is based on labeling single-stranded DNA fragments with a primer containing a 14 bp degenerate sequence. This allowed for an observed mutation frequency of ~0.001% mutations/bp in normal human genomic DNA [36]. Nevertheless, a number of highly sensitive genetic assays have indicated that the true mutation frequency in normal cells is likely to be far lower, with estimates of per-nucleotide mutation frequencies generally ranging from  $10^{-9}$  to  $10^{-11}$  [42]. Thus, the mutations seen in normal human genomic DNA by Kinde et al. are likely the result of significant technical artifacts.

**[0006]** Traditionally, next-generation sequencing platforms rely upon generation of sequence data from a single strand of DNA. As a consequence, artifactual mutations introduced during the initial rounds of PCR amplification are undetectable as errors - even with tagging techniques - if the base change is propagated to all subsequent PCR duplicates. Several types of DNA damage are highly mutagenic and may lead to this scenario. Spontaneous DNA damage arising from normal metabolic processes results in thousands of damaging events per cell per day [43]. In addition to damage from oxidative cellular processes, further DNA damage is generated *ex vivo* during tissue processing and DNA extraction [44]. These damage events can result in frequent

copying errors by DNA polymerases: for example a common DNA lesion arising from oxidative damage, 8-oxo-guanine, has the propensity to incorrectly pair with adenine during complementary strand extension with an overall efficiency greater than that of correct pairing with cytosine, and thus can contribute a large frequency of artifactual G→T mutations [45]. Likewise, deamination of cytosine to form uracil is a particularly common event which leads to the inappropriate insertion of adenine during PCR, thus producing artifactual C→T mutations with a frequency approaching 100% [46].

**[0007]** It would be desirable to develop an approach for tag-based error correction, which reduces or eliminates artifactual mutations arising from DNA damage, PCR errors, and sequencing errors; allows rare variants in heterogeneous populations to be detected with unprecedented sensitivity; and which capitalizes on the redundant information stored in complexed double-stranded DNA.

## **SUMMARY**

**[0008]** In one embodiment, a single molecule identifier (SMI) adaptor molecule for use in sequencing a double-stranded target nucleic acid molecule is provided. Said SMI adaptor molecule includes a double-stranded single molecule identifier (SMI) sequence which comprises a double-stranded degenerate or semi-degenerate DNA sequence; and an SMI ligation adaptor that allows the SMI adaptor molecule to be ligated to the double-stranded target nucleic acid sequence. In some embodiments, the double-stranded target nucleic acid molecule is a double-stranded DNA or RNA molecule.

**[0009]** In another embodiment, a method of obtaining the sequence of a double-stranded target nucleic acid is provided (also known as Duplex Consensus Sequencing

or DCS) is provided. Such a method may include steps of ligating a double-stranded target nucleic acid molecule to at least one SMI adaptor molecule to form a double-stranded SMI-target nucleic acid complex; amplifying the double-stranded SMI-target nucleic acid complex, resulting in a set of amplified SMI-target nucleic acid products; and sequencing the amplified SMI-target nucleic acid products.

**[0010]** In some embodiments, the method may additionally include generating an error-corrected double-stranded consensus sequence by (i) grouping the sequenced SMI-target nucleic acid products into families of paired target nucleic acid strands based on a common set of SMI sequences; and (ii) removing paired target nucleic acid strands having one or more nucleotide positions where the paired target nucleic acid strands are non-complementary (or alternatively removing individual nucleotide positions in cases where the sequence at the nucleotide position under consideration disagrees among the two strands). In further embodiments, the method confirms the presence of a true mutation by (i) identifying a mutation present in the paired target nucleic acid strands having one or more nucleotide positions that disagree; (ii) comparing the mutation present in the paired target nucleic acid strands to the error corrected double-stranded consensus sequence; and (iii) confirming the presence of a true mutation when the mutation is present on both of the target nucleic acid strands and appears in all members of a paired target nucleic acid family.

## **BRIEF DESCRIPTION OF THE DRAWINGS**

**[0011]** Figure 1 illustrates an overview of Duplex Consensus Sequencing. Sheared double-stranded DNA that has been end-repaired and T-tailed is combined with A-tailed SMI adaptors and ligated according to one embodiment. Because every

adaptor contains a unique, double-stranded, complementary n-mer random tag on each end (n-mer = 12 bp according to one embodiment), every DNA fragment becomes labeled with two distinct SMI sequences (arbitrarily designated  $\alpha$  and  $\beta$  in the single capture event shown). After size-selecting for appropriate length fragments, PCR amplification with primers containing Illumina flow-cell-compatible tails is carried out to generate families of PCR duplicates. By virtue of the asymmetric nature of adapted fragments, two types of PCR products are produced from each capture event. Those derived from one strand will have the  $\alpha$  SMI sequence adjacent to flow-cell sequence 1 and the  $\beta$  SMI sequence adjacent to flow cell sequence 2. PCR products originating from the complementary strand are labeled reciprocally.

**[0012]** Figure 2 illustrates Single Molecule Identifier (SMI) adaptor synthesis according to one embodiment. Oligonucleotides are annealed and the complement of the degenerate lower arm sequence (N's) plus adjacent fixed bases is produced by polymerase extension of the upper strand in the presence of all four dNTPs. After reaction cleanup, complete adaptor A-tailing is ensured by extended incubation with polymerase and dATP.

**[0013]** Figure 3 illustrates error correction through Duplex Consensus Sequencing (DCS) analysis according to one embodiment. (a-c) shows sequence reads (brown) sharing a unique set of SMI tags are grouped into paired families with members having strand identifiers in either the  $\alpha\beta$  or  $\beta\alpha$  orientation. Each family pair reflects one double-stranded DNA fragment. (a) shows mutations (spots) present in only one or a few family members representing sequencing mistakes or PCR-introduced errors occurring late in amplification. (b) shows mutations occurring in many or all

members of one family in a pair representing mutations scored on only one of the two strands, which can be due to PCR errors arising during the first round of amplification such as might occur when copying across sites of mutagenic DNA damage. (c) shows true mutations (\* arrow) present on both strands of a captured fragment appear in all members of a family pair. While artifactual mutations may co-occur in a family pair with a true mutation, these can be independently identified and discounted when producing (d) an error-corrected consensus sequence (+ arrow) for each duplex. (e) shows consensus sequences from all independently captured, randomly sheared fragments containing a particular genomic site are identified and (f) compared to determine the frequency of genetic variants at this locus within the sampled population.

**[0014]** Figure 4 illustrates an example of how a SMI sequence with n-mers of 4 nucleotides in length (4-mers) are read by Duplex Consensus Sequencing (DCS) according to some embodiments. (A) shows the 4-mers with the PCR primer binding sites (or flow cell sequences) 1 and 2 indicated at each end. (B) shows the same molecules as in (A) but with the strands separated and the lower strand now written in the 5'-3' direction. When these molecules are amplified with PCR and sequenced, they will yield the following sequence reads: The top strand will give a read 1 file of TAAC--- and a read 2 file of GCCA---. Combining the read 1 and read 2 tags will give TAACCGGA as the SMI for the top strand. The bottom strand will give a read 1 file of CGGA---- and a read 2 file of TAAC---. Combining the read 1 and read 2 tags will give CGGATAAC as the SMI for the bottom strand. (C) illustrates the orientation of paired strand mutations in DCS. In the initial DNA duplex shown in Figures 4A and 4B, a mutation "x" (which is paired to a complementary nucleotide "y") is shown on the left

side of the DNA duplex. The “x” will appear in read 1, and the complementary mutation on the opposite strand, “y,” will appear in read 2. Specifically, this would appear as “x” in both read 1 and read 2 data, because “y” in read 2 is read out as “x” by the sequencer owing to the nature of the sequencing primers, which generate the complementary sequence during read 2.

## **DETAILED DESCRIPTION**

**[0015]** Single molecule identifier adaptors and methods for their use are provided herein. According to the embodiments described herein, a single molecule identifier (SMI) adaptor molecule is provided. Said SMI adaptor molecule may include a double-stranded single molecule identifier (SMI) sequence, and an SMI ligation adaptor (Figure 2). Optionally, the SMI adaptor molecule further includes at least two PCR primer binding sites, at least two sequencing primer binding sites, or both.

**[0016]** In some embodiments, the SMI adaptor molecule includes a double-stranded, complementary SMI sequence (or “tag”) of nucleotides that is degenerate or semi-degenerate. In some embodiments, the degenerate or semi-degenerate SMI sequence may be a random degenerate sequence. The double-stranded SMI sequence includes a first degenerate or semi-degenerate nucleotide n-mer sequence and a second n-mer sequence that is complementary to the first degenerate or semi-degenerate nucleotide n-mer sequence. The first and second degenerate or semi-degenerate nucleotide n-mer sequences may be any suitable length to produce a sufficiently large number of unique tags to label a set of sheared DNA fragments from a segment of DNA. Each n-mer sequence may be between approximately 4 to 20 nucleotides in length. Therefore, each n-mer sequence may be approximately 4, 5, 6, 7,

8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 nucleotides in length. In one embodiment, the SMI sequence is a random degenerate nucleotide n-mer sequence which is 12 nucleotides in length. A 12 nucleotide SMI n-mer sequence that is ligated to each end of a target nucleic acid molecule, as described in the Example below, results in generation of up to  $4^{24}$  (i.e.,  $2.8 \times 10^{14}$ ) distinct tag sequences.

**[0017]** In some embodiments, the SMI tag nucleotide sequence may be completely random and degenerate, wherein each sequence position may be any nucleotide. (i.e., each position, represented by "X," is not limited, and may be an adenine (A), cytosine (C), guanine (G), thymine (T), or uracil (U)) or any other natural or non-natural DNA or RNA nucleotide or nucleotide-like substance or analog with base-pairing properties (e.g., xanthosine, inosine, hypoxanthine, xanthine, 7-methylguanine, 7-methylguanosine, 5,6-dihydrouracil, 5-methylcytosine, dihydouridine, isocytosine, isoguanine, deoxynucleosides, nucleosides, peptide nucleic acids, locked nucleic acids, glycol nucleic acids and threose nucleic acids). The term "nucleotide" as described herein, refers to any and all nucleotide or any suitable natural or non-natural DNA or RNA nucleotide or nucleotide-like substance or analog with base pairing properties as described above. In other embodiments, the sequences need not contain all possible bases at each position. The degenerate or semi-degenerate n-mer sequences may be generated by a polymerase-mediated method described in the Example below, or may be generated by preparing and annealing a library of individual oligonucleotides of known sequence. Alternatively, any degenerate or semi-degenerate n-mer sequences may be a randomly or non-randomly fragmented double stranded DNA molecule from any alternative source that differs from the target DNA source. In some embodiments,

the alternative source is a genome or plasmid derived from bacteria, an organism other than that of the target DNA, or a combination of such alternative organisms or sources. The random or non-random fragmented DNA may be introduced into SMI adaptors to serve as variable tags. This may be accomplished through enzymatic ligation or any other method known in the art.

**[0018]** In some embodiments, the SMI adaptor molecules are ligated to both ends of a target nucleic acid molecule, and then this complex is used according to the methods described below. In certain embodiments, it is not necessary to include n-mers on both adapter ends, however, it is more convenient because it means that one does not have to use two different types of adaptors and then select for ligated fragments that have one of each type rather than two of one type. The ability to determine which strand is which is still possible in the situation wherein only one of the two adaptors has a double-stranded SMI sequence.

**[0019]** In some embodiments, the SMI adaptor molecule may optionally include a double-stranded fixed reference sequence downstream of the n-mer sequences to help make ligation more uniform and help computationally filter out errors due to ligation problems with improperly synthesized adaptors. Each strand of the double-stranded fixed reference sequence may be 4 or 5 nucleotides in length sequence, however, the fixed reference sequence may be any suitable length including, but not limited to 3, 4, 5 or 6 nucleotides in length.

**[0020]** The SMI ligation adaptor may be any suitable ligation adaptor that is complementary to a ligation adaptor added to a double-stranded target nucleic acid sequence including, but not limited to a T-overhang, an A-overhang, a CG overhang, a

blunt end, or any other ligatable sequence. In some embodiments, the SMI ligation adaptor may be made using a method for A-tailing or T-tailing with polymerase extension; creating an overhang with a different enzyme; using a restriction enzyme to create a single or multiple nucleotide overhang, or any other method known in the art.

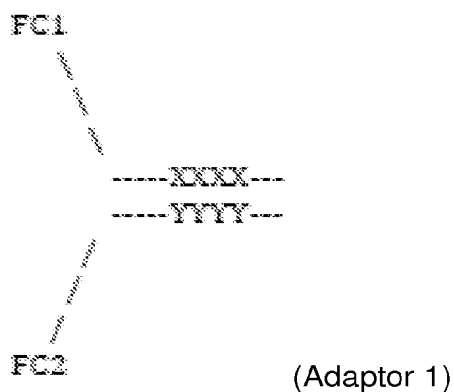
**[0021]** According to the embodiments described herein, the SMI adaptor molecule may include at least two PCR primer or “flow cell” binding sites: a forward PCR primer binding site (or a “flow cell 1” (FC1) binding site); and a reverse PCR primer binding site (or a “flow cell 2” (FC2) binding site). The SMI adaptor molecule may also include at least two sequencing primer binding sites, each corresponding to a sequencing read. Alternatively, the sequencing primer binding sites may be added in a separate step by inclusion of the necessary sequences as tails to the PCR primers, or by ligation of the needed sequences. Therefore, if a double-stranded target nucleic acid molecule has an SMI adaptor molecule ligated to each end, each sequenced strand will have two reads - a forward and a reverse read.

**[0022]** In some embodiments, the SMI adaptor molecule is a “Y-shaped” adaptor, which allows both strands to be independently amplified by a PCR method prior to sequencing because both the top and bottom strands have binding sites for PCR primers FC1 and FC2 as shown below. A schematic of a Y-shaped SMI adaptor molecule is also shown in Figure 2.

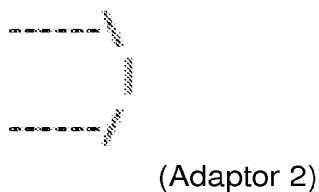
**[0023]** A Y-shaped SMI adaptor requires successful amplification and recovery of both strands. In one embodiment, a modification that would simplify consistent recovery of both strands entails ligation of a Y-shaped SMI adaptor molecule to one end of a DNA duplex molecule, and ligation of a “U-shaped” linker to the other end of the

molecule. PCR amplification of the hairpin-shaped product will then yield a linear fragment with flow cell sequences on either end. Distinct PCR primer binding sites (or flow cell sequences FC1 and FC2) will flank the DNA sequence corresponding to each of the two strands, and a given sequence seen in Read 1 will then have the sequence corresponding to the complementary DNA duplex strand seen in Read 2. Mutations can be scored only if they are seen on both ends of the molecule (corresponding to each strand of the original double-stranded fragment), i.e. at the same position in both Read 1 and Read 2. This design may be accomplished as follows.

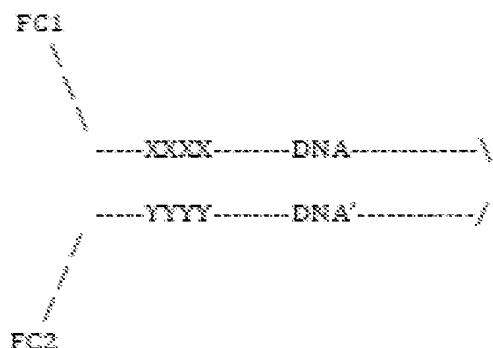
**[0024]** Adaptor 1 (shown below) is a Y-shaped SMI adaptor as described above (the SMI sequence is shown as X's in the top strand (a 4-mer), with the complementary bottom strand sequence shown as Y's):



**[0025]** Adaptor 2 (shown below) is a “U-shaped” linker:



**[0026]** Following ligation of both adaptors to a double-stranded target nucleic acid, the following structure is obtained:



**[0027]** When melted, the product will be of the following form (where “linker” is the sequence of adaptor 2):



**[0028]** This product is then PCR amplified. The reads will yield:

Read 1:

XXXX-----DNA-----

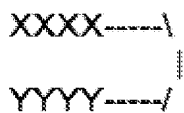
Read 2 (note that read 2 is seen as the complement of the bases sequenced:)

XXXX-----DNA-----

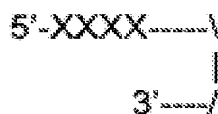
**[0029]** The sequences of the two duplex strands seen in the two sequence reads may then be compared, and sequence information and mutations will be scored only if the sequence at a given position matches in both of the reads.

**[0030]** This approach does not strictly require the use of an SMI tag, as the sheared ends can be used as identifiers to differentiate unique individual molecules from PCR duplicates. Thus the same concept would apply if one used any standard

sequencing adaptor as “Adaptor 1” and the U---shaped linker as “Adaptor 2.” However described below, there are a limited number of shear points flanking any given genomic position and thus the power to sequence deeply is increased via inclusion of the SMI tag. A hybrid method using a combination of sheared ends and a shorter n-mer tag (such as 1 or 2 or 3 or 4 or more degenerate or semi-degenerate bases) in the adaptor may also serve as unique molecular identifiers. Another design may include use of any sequencing adaptor (such as one lacking an n-mer tag) in conjunction with an n-mer tag that is incorporated into the U-shaped linker molecule. Such a design would be of the following form (where X and Y represent complementary degenerate or semi-degenerate nucleotides):



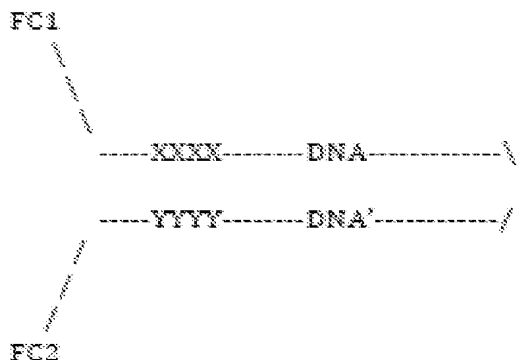
**[0031]** Synthesis of such a design may be obtained in a number of ways, for example synthesizing a set of hairpin oligonucleotides in which each individual oligonucleotide encodes a complementary n-mer sequence, or alternatively by using a DNA polymerase to carry out extension from the following product (where X’s represent degenerate nucleotides):



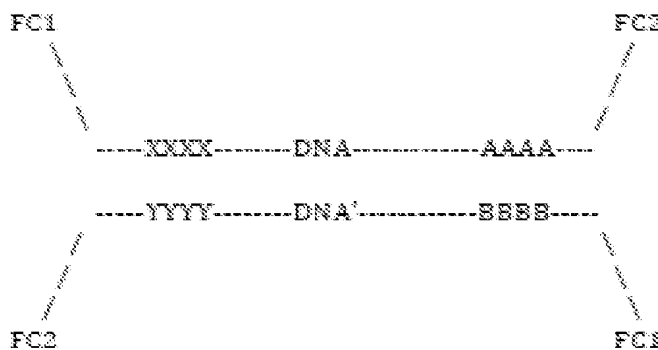
**[0032]** Inclusion of the SMI tag is also extremely useful for identifying correct ligation products, as the assay uses two distinct adaptors. This will yield multiple

possible ligation products:

[0033] **Product I.** Adaptor 1-----DNA-----Adaptor 2, which yields the desired product:

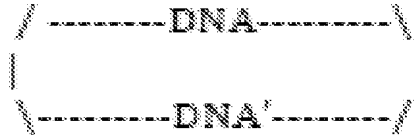


[0034] **Product II.** Adaptor 1-----DNA-----Adaptor 1. This will result in the DNA being amplified as two separate strands, i.e. as occurs in the DCS approach described elsewhere in this document (the second copy of Adaptor 1 is shown below with the SMI as AAA – BBB to emphasize that every DCS adaptor has a distinct SMI sequence)



[0035] **Product III.** Adaptor 2-----DNA-----Adaptor 2. This will result in a

non-amplifiable circular product shown below:



**[0036]** Product III is non-amplifiable, given the absence of primer binding sites and thus will not be present in the final DNA sequences. Thus only Product II needs to be avoided. The formation of Product II can be minimized in the ligation step by using an excess of Adaptor 2 (relative to Adaptor 1). Then primarily Products I and III will be obtained, with minimal formation of Product II. Additionally, a variety of biochemical means of enriching for products containing adaptor 2 are possible such as using affinity probes that are complementary to the hairpin loop sequence itself. Product I results in the same SMI sequence in both the Read 1 and Read 2 sequence reads. In the example depicted above, Product I sequences can thus be identified by virtue of having matching SMIs of the form XXXX in Read 1 and XXXX in Read 2.

**[0037]** By contrast, in the case of Product II, the SMI sequences on either end of the sequenced molecule will arise from distinct DCS adaptors having different SMI sequences. In the example shown above, Product II sequences yield SMIs of the form XXXX (Read 1) – BBB (Read 2) upon sequencing of the top strand, and BBBB (Read 1) – XXXX (Read 2) upon sequencing of the bottom strand. Thus Product II sequences can be easily identified and computationally removed from the final sequence data.

**[0038]** Data resulting from Product II is useful, because Product II corresponds to the product analyzed under the approach detailed in the Example below. Product I contains a self-complementary hairpin sequence that can impair polymerase extension

during amplification, however, this type of amplification has already been enabled in the technique of “Hairpin PCR” [50] which involves linking of the two strands followed by amplification with gene-specific primers. Amplification conditions that are compatible with amplification of hairpin DNA are thus already established. Moreover, ligation and amplification with circularizing “linkers” (i.e. hairpin adaptors affixed to both ends of a fragment) has been demonstrated as a step in the Pacific Biosciences sample preparation workflow [49]. As the sequence of the linker itself does not matter in the workflow, the published linker sequences from either of these references would be adequate for use in the assay.

**[0039]** In some aspects of some embodiments, deliberate ligation of “U-shaped” adaptors containing 1) a double-stranded n-mer (or other form of degenerate or semi-degenerate double-stranded tag as enumerated above) plus 2) primer binding sites to both ends of a captured fragment may be desirable. Producing closed circles of captured material may help facilitate removal of non-captured DNA by exonuclease digestion given that circularized DNA will be protected from digestion by such enzymes. Additionally, closed circles may be pre-amplified using rolling circle amplification or serve as the substrate for continuous loop sequencing [49]. Recognition sites for restriction endonuclease digestion could be engineered into these adaptors to render closed loops open once again if more convenient for subsequent steps.

**[0040]** The SMI adaptor molecules described herein have several uses. In some embodiments, the SMI adaptor molecules described herein may be used in methods to obtain the sequence or other sequence-related information of a double-stranded target nucleic acid molecule. According to the embodiments described herein, the term

“double-stranded target nucleic acid molecule” includes a double-stranded DNA molecule or a double-stranded RNA molecule. Thus, the SMI adaptor molecules and methods of use described herein are applicable to genotyping and other applications related to sequencing of DNA molecules, but are also applicable to RNA sequencing applications such as for sequencing of double-stranded RNA viruses. Methods for sequencing RNA may include any of the embodiments described herein with respect to DNA sequencing, and vice-versa. For example, any double stranded target nucleic acid molecule may be ligated to an SMI adaptor molecule which includes a double-stranded RNA or DNA n-mer tag and an RNA or DNA ligation adaptor as described above. Methods exist for directly sequencing RNA [51]; alternatively, the ligated product may be reverse transcribed into DNA, and then sequenced as a double-stranded target DNA molecule.

**[0041]** In one embodiment, the double-stranded target nucleic acid molecule may be a sheared double-stranded DNA or RNA fragment. The sheared target DNA or RNA molecule may be end repaired and a double-stranded target nucleic acid sequence ligation adaptor may be added to each end of the sheared target DNA or RNA molecule. The double-stranded target nucleic acid sequence ligation adaptor may be any suitable ligation adaptor that is complementary to the SMI ligation adaptor described above including, but not limited to a T-overhang, an A-overhang, a CG overhang, blunt end or any other ligatable sequence. In some embodiments, the double-stranded target nucleic acid sequence ligation adaptor may be made using a method for A-tailing or T-tailing with polymerase extension; adding an overhang with a different enzyme; using a restriction enzyme to create a ligatable overhang; or any other method known in the art.

**[0042]** Methods to obtain the sequence or other sequence-related information of a double-stranded target nucleic acid molecule may include a step of ligating the double-stranded target nucleic acid molecule to at least one SMI adaptor molecule, such as those described above, to form a double-stranded target nucleic acid complex. In one embodiment, each end of the double-stranded target nucleic acid molecule is ligated to an SMI adaptor molecule. The double-stranded target nucleic acid complex is then amplified by a method known in the art (e.g., a PCR or non-PCR method known in the art), resulting in a set of uniquely labeled, amplified SMI-target nucleic acid products. These products are then sequenced using any suitable method known in the art including, but not limited to, the Illumina sequencing platform, ABI SOLiD sequencing platform, Pacific Biosciences sequencing platform, 454 Life Sciences sequencing platform, Ion Torrent sequencing platform, Helicos sequencing platform, and nanopore sequencing technology.

**[0043]** In certain embodiments, a method of generating an error corrected double-stranded consensus sequence is provided. In such a method, the sequenced SMI-DNA products generated in the methods described above are grouped into families of paired target nucleic acid strands based on a common set of SMI sequences. Then, the paired target nucleic acid strands can be filtered to remove nucleotide positions where the sequences seen on both of the paired partner DNA strands are not complementary. This error corrected double-stranded consensus sequence may be used in a method for confirming the presence of a true mutation (as opposed to a PCR error or other artifactual mutation) in a target nucleic acid sequence. According to certain embodiments, such a method may include identifying one or more mutations

present in the paired target nucleic acid strands that have one or more nucleotide positions that disagree between the two strands, then comparing the mutation present in the paired target nucleic acid strands to the error corrected double-stranded consensus sequence. The presence of a true mutation is confirmed when the mutation is present on both of the target nucleic acid strands and also appear in all members of a paired target nucleic acid family.

**[0044]** The accuracy of current approaches to next-generation sequencing is limited due to their dependence on interrogating single-stranded DNA. This dependence makes potential sources of error such as PCR amplification errors and DNA damage fundamentally limiting. However, the complementary strands of a double-stranded DNA molecule (or "DNA duplex") contain redundant sequencing information (i.e., one molecule reciprocally encoding the sequence information of its partner) which can be utilized to eliminate such artifacts. Limitations related to sequencing single-stranded DNA (e.g., sequencing errors) may therefore be overcome using the methods described herein. This is accomplished by individually tagging and sequencing each of the two strands of a double-stranded (or duplex) target nucleic acid molecule and comparing the individual tagged amplicons derived from one half of a double-stranded complex with those of the other half of the same molecule. Duplex Consensus Sequencing (DCS), significantly lowers the error rate of sequencing. In some embodiments, the DCS method may be used in methods for high sensitivity detection of rare mutant and variant DNA as described further below.

**[0045]** As described above, one approach that has previously been reported for DNA sequencing involves incorporation of a random tag sequence into a PCR primer

[36]. This approach results in an improvement in accuracy relative to standard Illumina sequencing, but is fundamentally limited in that it is based upon amplification and sequencing of single-stranded DNA and thus cannot overcome limitations in sensitivity owing to single-stranded DNA damage events. In the methods described herein, PCR duplicates are generated from a single strand of DNA, and the sequences of the duplicates are compared. Mutations are scored only when they are present in multiple replicates of a single starting molecule. The DCS approach overcomes the limitation of previous approaches by considering both DNA strands.

**[0046]** DNA damage should not be a limiting factor in DCS, because miscoding damage events at a single base-pair position occur essentially exclusively on only one of the two DNA strands. For DNA damage to result in an artifactual mutation in DCS, damage would need to be present at the same nucleotide position on both strands. Even if complementary nucleotides in a duplex were both damaged, the damage would need to result in complementary sequencing errors to result in mis-scoring of a mutation. Likewise, spontaneous PCR errors would need to result in complementary mutations at the same position on both strands; with a first-round mutation frequency of Taq polymerase of approximately  $10^{-5}$  and three possible incorrect bases that could be mis-inserted, the probability of two complementary PCR errors occurring would be  $10^{-5} \times 10^{-5} \times 1/3 = 3.3 \times 10^{-11}$

**[0047]** According to some embodiments, the sequencing method may be performed using the Illumina or similar platforms including those enumerated above without the use of SMI adaptor molecules, but instead by using the random shear points of DNA as identifiers. For a given DNA sequence seen in sequencing read 1 with a

specific set of shear points, the partner strand will be seen as a matching sequence in read two with identical shear points. In practice, this approach is limited by the limited number of possible shear points that overlap any given DNA position. However, according to some embodiments, shear points of a target nucleic acid molecule may be used as unique identifiers to identify double-stranded (or duplex) pairs, resulting in an apparent error frequency at least as low as that seen with traditional sequencing methods, but with a significantly lower loss of sequence capacity. In other embodiments, DCS based on shear points alone may have a role for confirmation that specific mutations of interest are true mutations which were indeed present in the starting sample (i.e. present in both DNA strands), as opposed to being PCR or sequencing artifacts. Overall, however, DCS is most generally applicable when randomized, complementary double-stranded SMI sequences are used. A 24 nucleotide double-stranded SMI sequence was used in the Example described below, which may yield up to  $4^{24} = 2.8 \times 10^{14}$  distinct double-stranded SMI sequences. Combining information regarding the shear points of DNA with the SMI tag sequence would allow a shorter SMI to be used, thus minimizing loss of sequencing capacity due to sequencing of the SMI itself.

**[0048]** In certain embodiments, the SMI adaptor molecules may also be used in methods of single-molecule counting for accurate determination of DNA or RNA copy number [38]. Again, since the SMI tags are present in the adaptors, there are no altered steps required in library preparation, which is in contrast to other methods for using random tags for single-molecule counting. Single-molecule counting has a large number of applications including, but not limited to, accurate detection of altered

genomic copy number (e.g., for sensitive diagnosis of genetic conditions such as trisomy 21 [47]), for accurate identification of altered mRNA copy number in transcriptional sequencing and chromatin immunoprecipitation experiments, quantification of circulating microRNAs, quantification of viral load of DNA or RNA viruses, quantification of microorganism abundance, quantification of circulating neoplastic cells, counting of DNA-labeled molecules of any variety including tagged antibodies or aptamers, and quantification of relative abundances of different individual's genomes in forensic applications.

**[0049]** In another embodiment, the SMI adaptor molecules may be used in methods for unambiguous identification of PCR duplicates. In order to restrict sequencing analysis to uniquely sequenced DNA fragments, many sequencing studies include a step to filter out PCR duplicates by using the shear points at the ends of DNA molecules to identify distinct molecules. When multiple molecules exhibit identical shear points, all but one of the molecules are discarded from analysis under the assumption that the molecules represent multiple PCR copies of the same starting molecule. However sequence reads with identical shear points can also reflect distinct molecules because there are a limited number of possible shear points at any given genomic location, and with increasing sequencing depth, recurrent shear points are increasingly likely to be seen [48]. Because the use of SMI tags (or “double-stranded SMI sequences”) allows every molecule to be uniquely labeled prior to PCR duplication, true PCR duplicates may be unambiguously identified by virtue of having a common (i.e., the same or identical) SMI sequence. This approach would thereby minimize the loss of data by overcoming the intrinsic limitations of using shear points to identify PCR

duplicates.

**[0050]** Importantly, once SMI-containing adaptors are synthesized by a straightforward series of enzymatic steps or are produced through synthesis of a set of oligonucleotides containing complementary tag sequences, they may be substituted for standard sequencing adaptors. Thus, use of DCS does not require any significant deviations from the normal workflow of sample preparation for Illumina DNA sequencing. Moreover, the DCS approach can be generalized to nearly any sequencing platform because a double-stranded SMI tag can be incorporated into other existing adaptors, or for sequencing approaches that do not require adaptors, a double-stranded SMI tag can be ligated onto duplex DNA sample prior to sequencing. The compatibility of DCS with existing sequencing workflows, the potential for greatly reducing the error rate of DNA sequencing, and the multitude of applications for the double-stranded SMI sequences validate DCS as a technique that may play a general role in next generation DNA sequencing.

**[0051]** The following examples are intended to illustrate various embodiments of the invention. As such, the specific embodiments discussed are not to be construed as limitations on the scope of the invention. It will be apparent to one skilled in the art that various equivalents, changes, and modifications may be made without departing from the scope of invention, and it is understood that such equivalent embodiments are to be included herein. Further, all references cited in the disclosure are hereby incorporated by reference in their entirety, as if fully set forth herein.

## EXAMPLES

### **Example 1: Generation of SMI Adaptor Molecules and their use in sequencing double-stranded target DNA**

#### Materials and Methods

**[0052]** *Materials.* Oligonucleotides were from IDT and were ordered as PAGE purified. Klenow exo- was from NEB. T4 ligase was from Enzymatics.

**[0053]** *DNA isolation.* Genomic DNA was isolated from normal human colonic mucosa by sodium iodide extraction (Wako Chemicals USA).

**[0054]** *Adaptor synthesis.* The adaptors were synthesized from two oligos, designated as:

the primer strand:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA  
CGCTCTTCCGATCT (SEQ ID NO:1); and

the template strand:

/5phos/ACTGNNNNNNNNNNNAGATCGGAAGAGCACACGTCTG  
AACTCCAGTCAC (SEQ ID NO:2).

**[0055]** The two adaptor strands were annealed by combining equimolar amounts of each oligo to a final concentration of 50 micromolar and heating to 95°C for 5 minutes. The oligo mix was allowed to cool to room temperature for over 1 hour. The annealed primer-template complex was extended in a reaction consisting of 40 micromolar primer-template, 25 units Klenow exo-, 250 micromolar each dNTP, 50 mM NaCl, 10

mM Tris-HCl pH 7.9, 10 mM MgCl<sub>2</sub>, and 1 mM dithiothreitol. The product was isolated by ethanol precipitation, and was then A-tailed with 25 units Klenow exo-, 1 mM dATP, 50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM MgCl<sub>2</sub>, and 1 mM dithiothreitol. The product was again ethanol precipitated and resuspended to a final concentration of 50 micromolar.

**[0056]**      *Sequencing library preparation.* 3 micrograms of DNA was diluted into 130 microliters of TE buffer (10 mM tris-HCl, pH 8.0, 0.1 M EDTA) and was sheared on the Covaris AFA system with duty cycle 10%, intensity 5, cycles/burst 200, time 20 seconds x 6, temperature 4 C. DNA was purified with 2 volumes of Agencourt AMPure XP beads per the manufacturer's protocol. After end-repair with the NEB end-repair kit per the manufacturer's protocol, DNA fragments larger than the optimal range of ~200-500 bp were removed by adding 0.7 volumes of AMPure XP beads and transferring the supernatant to a separate tube (fragments larger than 500 bp bind to the beads and are discarded). An additional 0.65 volumes of AMPure XP beads were added (this step allows fragments of approximately 200bp or greater to bind to the beads). The beads were washed and DNA eluted. DNA was then T-tailed in a reaction containing 5 units Klenow exo-, 1 mM dTTP, 50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM MgCl<sub>2</sub>, 1 mM. The reaction proceeded for 1 hour at 37 C. DNA was purified with 1.2 volumes of AMPure XP beads. The custom adaptors were ligated by combining 750 ng of T-tailed DNA with 250 pmol adaptors in a reaction containing 3000 units T4 DNA ligase, 50 mM Tris-HCl pH 7.6, 10 mM MgCl<sub>2</sub>, 5 mM DTT, 1 mM ATP. The reaction was incubated 25 C for 15 minutes, and purified with 1.2 volumes of AMPure XP beads.

**[0057]**      *Pre-capture amplification.* 375 ng adaptor-ligated DNA was PCR amplified

with primers AATGATACGGCGACCACCGAG (SEQ ID NO:3) and GTGACTGGAGTTCAGACGTGTGC (SEQ ID NO:4) using the Kappa high-fidelity PCR kit for 8 cycles with an annealing temperature of 60 C. The product was purified with 1.2 volumes of AMPure XP beads.

**[0058]** *DNA capture.* Target capture was performed with the Agilent SureSelect system per the manufacturer's recommendations, except that capture volumes were performed at one-half of the standard volume. The capture set targeted an arbitrary 758 kb region of the genome consisting of both coding and noncoding sequences. Capture baits were 120 nt in length, and were prepared with the Agilent eArray tool with 3x tiling.

**[0059]** *Post-capture amplification.* Captured DNA was amplified with PCR primers AATGATACGGCGACCACCGAG (SEQ ID NO:3) and CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGC (SEQ ID NO:5) where XXXXXX indicates the position of a fixed multiplexing barcode sequence). 2.0 fmol of DNA was used per lane for sequencing on an Illumina HiSeq 2000.

**[0060]** *Data processing.* Reads with intact SMI adaptors include a 12 nucleotide random sequence, followed by a 5 nucleotide fixed sequence. These reads were identified by filtering out reads that lack the expected fixed sequence at positions 13-17. The SMI sequence from both the forward and reverse sequencing reads (i.e., the first and second degenerate n-mer sequences) was computationally added to the read header, and the fixed sequence removed. The first 4 nucleotides located following the adaptor sequence were also removed due to the propensity for ligation and end-repair errors to result in an elevated error rate near the end of the DNA fragments. Reads

having common (i.e., identical) SMI sequences were grouped together, and were collapsed to generate a consensus read. Sequencing positions were discounted if the consensus group covering that position consisted of fewer than 3 members, or if fewer than 90% of the sequences at that position in the consensus group had the identical sequence. Reads were aligned to the human genome with the Burrows-Wheeler Aligner. The consensus sequences were then paired with their strand-mate by grouping each 24 nucleotide tag of form AB in read 1 with its corresponding tag of form BA in read 2. Resultant sequence positions were considered only when information from both DNA strands was in perfect agreement. An overview of the data processing workflow is as follows:

1. Discard reads that do not have the 5 nt fixed reference (or “spacer”) sequence (CAGTA; SEQ ID NO:6) present after 12 random nucleotides.
2. Combine the 12 nt SMI tags from read 1 and read 2, and transfer the combined 24 nt SMI sequence into the read header.
3. Discard SMIs with inadequate complexity (i.e., those with > 10 consecutive identical nucleotides).
4. Remove the 5 nt fixed reference sequence.
5. Trim an additional 4 nt from the 5’ ends of each read pair (sites of error prone end repair).
6. Group together reads which have identical 24 nt SMIs.
7. Collapse to SMI consensus reads, scoring only positions with 3 or more SMI duplicates and >90% sequence identity among the duplicates.
8. For each read in read 1 file having SMI of format AB, group with corresponding DCS partner in read 2 with SMI of format BA.
9. Only score positions with identical sequence among both DCS partners.
10. Align reads to the human genome.

**[0061]** Code for carrying out the workflow may be pre-existing or may involve programming within the skill of those in the art. In some embodiments, however, the

Python code, which is illustrated in Appendix A, may be used for carrying out the pairing and scoring of partner strands according to steps 8 and 9 of the workflow described above. (Appendix A, is hereby incorporated by reference in its entirety as if fully set forth herein).

### Results

**[0062]** To overcome limitations in the sensitivity of variant detection by single-stranded next-generation DNA sequencing, an alternative approach to library preparation and analysis was designed, which is known herein as Duplex Consensus Sequencing (DCS) (Figure 1). The DCS method described herein involves tagging both strands of duplex DNA with a random, yet complementary double-stranded nucleotide sequence, which is known herein as a double-stranded single molecule identifier (SMI) sequence. The individually tagged strands are then PCR amplified. Every duplicate that arises from a single strand of DNA will have the same SMI, and thus each strand in a DNA duplex pair generates a distinct, yet related population of PCR duplicates after amplification owing to the complementary nature of the SMIs on the two strands of the duplex. Comparing the sequence obtained from each of the two strands comprising a single molecule of duplex DNA facilitates differentiation of sequencing errors from true mutations. When an apparent mutation is, due to a PCR or sequencing error, the substitution will only be seen on a single strand. In contrast, with a true DNA mutation, complementary substitutions will be present on both strands (see Figure 4C).

**[0063]** Following tagging with a double-stranded SMI and PCR amplification, a family of molecules is obtained that arose from a single DNA molecule; members of the same PCR “family” are then grouped together by virtue of having a common (i.e., the

same) SMI tag sequence. The sequences of uniquely tagged PCR duplicates are subsequently compared in order to create a PCR consensus sequence. Only DNA positions that yield the same DNA sequence in a specified proportion of the PCR duplicates in a family, such as 90% of the duplicates in one embodiment, are used to create the PCR consensus sequence. Next, PCR consensus sequences arising from two complementary strands of duplex DNA can be identified by virtue of the complementary SMIs (Figure 3) to identify the “partner SMI.” Specifically, a 24-nucleotide SMI consists of two 12-nucleotide sequences that can be designated XY. For an SMI of form XY in read 1, the partner SMI will be of form YX in read 2. An example to illustrate this point is given in Figure 4. Following partnering of two strands by virtue of their complementary SMIs, the sequences of the strands are compared. Sequence reads at a given position are kept only if the read data from each of the two paired strands is in agreement.

**[0064]** In order to label each of the strands of duplex DNA with unique complementary tags, adaptors which contain the standard sequences required for the Illumina HiSeq system were synthesized, but with addition of a double-stranded, complementary SMI sequence (or “tag”) of 12 random nucleotides (or a random “degenerate sequence”) per strand. Target DNA molecules having a random SMI sequence n-mer that is 12 nucleotides in length on each end will therefore have a unique 24 nucleotide SMI sequence. The adaptors were prepared (Figure 2) from two partially complementary oligonucleotides, one of which has a single-stranded 12 nucleotide random nucleotide sequence (i.e. a first random degenerate nucleotide n-mer sequence) followed by a single stranded fixed reference sequence that is 4

nucleotides in length. The single-stranded random nucleotide tag was converted to a double-stranded, complementary SMI tag by extension with Klenow exo- DNA polymerase and the extended adaptor was purified by ethanol precipitation. Due to the partial A-tailing property of Klenow exo-, this protocol results in a mixture of blunt-ended adaptors and adaptors with a single-nucleotide A overhang (data not shown). A single-nucleotide A-overhang was added to the residual blunt fragments by incubating the adaptors with Klenow exo- DNA polymerase and a high concentration of dATP (1 mM), and purified the adaptors again by ethanol precipitation.

**[0065]** DNA for sequencing was sheared and end-repaired by standard methods, with size-selection for fragments in the range of ~200-500 bp by size-selective binding to Ampure XP beads. Standard Illumina library preparation protocols involve ligating A-tailed DNA to T-tailed adaptors. However, because A-tailed adaptors were used, the DNA was T-tailed by incubating the end-repaired DNA with Klenow exo- DNA polymerase and 1 mM dTTP. The adaptor-ligated library was PCR amplified and subjected to SureSelect capture, with targeting of an arbitrary 758 kb portion of the genome (DNA coordinates available upon request). The efficiency of adaptor ligation, PCR amplification, DNA capture, and sequencing were comparable to those seen with standard library preparation methods (data not shown). Although Agilent Sure Select probes are used in this example, any suitable method of DNA selection may be used to capture particular target double-stranded DNA sequences. For example, selection and capture may be accomplished by any selection by hybridization method (e.g., Agilent SureSelect, Primer Extension Capture, exploitation of biotinylated PCR amplicons as bait, Agilent HaloPlex) wherein probes that target the desired double-stranded DNA

sequence may be recovered by an in-array capture (using probes immobilized on glass slides) or by affinity using magnetic beads in an in-solution capture. In addition, mitochondrial and some other forms of DNA may be isolated by size selection. Alternatively, in some embodiments, no enrichment is performed.

**[0066]** This protocol was used to sequence DNA isolated from normal colonic mucosa. Mutations were initially scored without consideration of the SMI sequences. PCR duplicates were filtered out with samtools rmdup, a standard tool which uses the shear points of DNA molecules to identify PCR duplicates, as molecules arising from duplicated DNA will have shared shear points. In order to focus specifically on non-clonal mutations, only those positions in the genome with at least 20x coverage and at which fewer than 5% of reads differed from the hg19 reference sequence were considered. This approach resulted in 70.9 million nucleotides of sequence data and 56,890 mutations, indicating an overall mutation frequency of  $8.03 \times 10^{-4}$ , in accord with the error rate of Illumina next-generation sequencing of ~0.1-1% [32].

**[0067]** Next, the SMI tags were used to group together PCR duplicates that arose from individual single-stranded DNA molecules and to create a consensus sequence from the family of duplicates. At least 3 PCR duplicates were required, with at least 90% agreement in sequence among all duplicates, to consider a site for mutations. Scoring the mutation frequency as above, again considering only sites with a minimum of 20x coverage and with <5% of reads differing from reference, resulted in 145 million nucleotides of sequence with 6,508 mutations and an overall mutation frequency of  $4.47 \times 10^{-5}$ , consistent with prior reports [36]. Notably, far more nucleotides of DNA sequence were obtained in this approach (145 million) than in the standard Illumina

sequencing approach (70 million) detailed above which is dependent on use of the shear points of single-ended reads to identify PCR duplicates. The improved sequence coverage arose from use of the SMI to identify PCR duplicates, because identifying PCR duplicates by consideration of uniquely sheared DNA ends is fundamentally limited by the small number of possible shear points that overlap a given position of the genome and the propensity for specific genomic regions to be more readily undergo shearing. Thus filtering PCR duplicates by using shear points resulted in discarding a large portion of the reads.

**[0068]** Finally, the complementary nature of the double-stranded SMI sequences was used to identify pairs of consensus groups that arose from complementary DNA strands. Sequence reads were considered only when the read data from each of the two strands is in perfect agreement. In a pilot experiment, after grouping of PCR duplicates as above, 29,409 SMI partner pairs were found, indicative that fewer than 1% of tags had their corresponding partner tag present in the library. The low recovery of tag pairs was most likely due to inadequate amplification of the starting DNA library. Among these tag-pairs, 24,772 duplex consensus strands were identified with an average strand length of 82 nucleotides, resulting in 2 million nucleotides of DNA consensus sequence. The sequences of the paired duplex strands disagreed at 3,585 of the nucleotide positions, indicative of single-stranded errors (i.e. PCR or sequencing errors); these sites of disagreement were removed, leaving only bases at which the sequence of both duplex strands were in perfect agreement. Next, as above, analysis of mutation frequencies was restricted to sites with at least 10x coverage and at which fewer than 10% of reads disagreed from the hg19 reference sequence. Because the 2

million nucleotides of read data were spread across a 758kb target, our average depth was only ~3x. Thus only 14,464 nucleotides of DNA sequence corresponded to sites with at least 10x depth. Among these sites, zero mutations were seen. To increase the number of tag pairs considered, analysis described above was repeated, but PCR duplicates were grouped with a minimum of only 1 duplicate per site. This resulted in 28,359 nucleotides of DNA sequence with at least 10x depth. Again, no mutations were detected.

**[0069]** Current experiments are being performed on vastly smaller target DNA molecules (ranging from ~300 bp to ~20 kb in size). Use of smaller DNA targets will allow for much greater sequencing depth, and far more accurate assessment of the background mutation rate of the assay. In addition, the protocol has been modified to incorporate a greater number of PCR cycles initiated off a smaller number of genome equivalents, which will increase the fraction of tags for which both of the partner tag strands have been sufficiently amplified to be represented in the final sequence data. Indeed, among the 3.6 million SMIs present in our initial library which underwent PCR duplication, 1.5 million of the SMIs were present only once, indicating insufficient amplification of the DNA due in part to the low number of PCR cycles used.

**Example 2: Demonstration of error-correction by DCS using randomly sheared  
DNA ends as Single Molecule Identifiers**

Methods

**[0070]** *Sequencing library preparation.* Genomic DNA was isolated from a derivative of *Saccharomyces cerevisiae* strain SC288 by standard methods. The DNA

was randomly sheared by the Covaris AFA system, followed by end-repair, A-tailing, and ligation of Illumina TruSeq DNA sequencing adaptors, all by standard library preparation methods. The resultant sequence data consisted of an average 32.5 fold depth of the 12 megabase *S. cerevisiae* genome.

**[0071]** *Data analysis.* The first 10 nucleotides of each sequencing read pair, corresponding to the randomly sheared DNA ends, were combined, such that the first 10 nucleotides of read 1, referred to as A, was combined with the first 10 nucleotides of read 2, referred to as B, to yield an SMI tag of form AB. Reads were grouped according to SMI sequence, and nucleotide reads were considered only if they agreed among at least 90% of family members sharing a given tag sequence. For DCS analysis, a tag of form AB1 is partnered with the corresponding tag of form BA2, and nucleotide positions are considered only when the sequence is in agreement among read pairs with matching tags AB1 and BA2.

### Results

**[0072]** In order to demonstrate the capability of DCS analysis to remove sequencing errors, a sequencing library was prepared under standard conditions with commercially available sequencing adaptors, and the randomly sheared DNA ends were used as SMI's. First, reads were grouped by SMI with a minimum family size of 1 member. Considering only sites with a minimum of 20x coverage and with <5% of reads differing from reference, this analysis resulted in 644.8 million nucleotides of sequence data and 2,381,428 mutations, yielding an overall mutation frequency of  $3.69 \times 10^{-3}$ .

**[0073]** The data was then subjected to DCS analysis with the SMI tags,

searching for tags of form AB1 that have partner tags of form BA2, and considered only positions at which the sequence from the two strands was in perfect agreement. 3.1% of the tags had a matching partner present in the library, resulting in 2.9 million nucleotides of sequence data. The sequences of the duplex strands were not complementary at 40,874 nucleotide positions; these disagreeing positions, representing likely sequencing or PCR errors, were removed from analysis. Again considering positions with at least 20x coverage and <5% of reads differing from reference, 3.0 million nucleotides of sequence data and 157 mutations were obtained, with an overall mutation frequency of  $5.33 \times 10^{-5}$ , indicative of removal of >98% of mutations seen in raw analysis and thereby demonstrating the capability of DCS to lower the error rate of DNA sequencing.

**[0074]** To compare this result to the method of Kinde et al. [36], reads were grouped into families by SMI tag as before but filtered for families with a minimum of 3 members. This resulted in 1.4 million nucleotides of sequence data and 61 mutations, with an overall mutation frequency of  $4.25 \times 10^{-5}$ . Thus, the method of Kinde et al., with a minimum family size of 3, resulted in less than half as much resultant sequence data after filtering than was obtained by DCS with a minimum family size of 1. Thus, DCS lowered the error rate of sequencing to a comparable degree to a method considered state-of-the-art, but with less loss of sequencing capacity.

### Discussion

**[0075]** It was demonstrated that DCS analysis, using sheared DNA ends as unique molecular identifiers, results in a lowering of the apparent error rate of DNA sequencing. As this proof-of-concept experiment was performed on a library that was

not optimized to maximize recovery of both strands, there were not sufficient strand-pairs recovered to perform DCS analysis with a minimum family size of >1 member. Requiring family sizes >1 is expected to further reduce sequencing errors. Moreover, this analysis was limited in that it did not include ligation of degenerate SMI tag sequences; owing to the limited number of shear points flanking any given nucleotide position, use of shear points as SMIs limits the number of unique molecules that can be sequenced in a single experiment. The use of shear points as SMIs in conjunction with an exogenously ligated SMI tag sequence would allow for increased depth of sequencing at any given nucleotide position.

## REFERENCES

The references, patents and published patent applications listed below, and all references cited in the specification above are hereby incorporated by reference in their entirety, as if fully set forth herein.

- [1] Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2010;11:31-46.
- [2] Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008;26:1135-45.
- [3] Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, Baerlocher L, et al. Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc Natl Acad Sci USA.* 2011;108:13177-82.
- [4] Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature.* 2011;480:368-71.
- [5] García-Garcerà M, Gigli E, Sanchez-Quinto F, Ramirez O, Calafell F, Civit S, et al. Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics. *PLoS ONE.* 2011;6:e24161.
- [6] Fordyce SL, Ávila-Arcos MC, Rockenbauer E, Børsting C, Frank-Hansen R, Petersen FT, et al. High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *BioTechniques.* 2011;51:127-33.
- [7] Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, et al. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods.* 2009;6:263-5.
- [8] Out AA, van Minderhout IJHM, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, et al. Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat.* 2009;30:1703-12.
- [9] Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA.* 2008;105:16266-71.
- [10] Chiu RWK, Akolekar R, Zheng YWL, Leung TY, Sun H, Chan KCA, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *BMJ.* 2011;342:c7401.

- [11] Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci USA*. 2008;105:10513-8.
- [12] Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*. 2015;481:506-9.
- [13] Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Science Translational Medicine*. 2009;1:12ra23-12ra23.
- [14] Hyman RW, Herndon CN, Jiang H, Palm C, Fukushima M, Bernstein D, et al. The dynamics of the vaginal microbiome during infertility therapy with in vitro fertilization-embryo transfer. *J Assist Reprod Genet*. 2012;29:105-15.
- [15] LaTuga MS, Ellis JC, Cotton CM, Goldberg RN, Wynn JL, Jackson RB, et al. Beyond bacteria: a study of the enteric microbial consortium in extremely low birth weight infants. *PLoS ONE*. 2011;6:e27858.
- [16] Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: interindividual variation and dynamic response to diet. *Genome Res*. 2011;21:1616-25.
- [17] Yang J, Yang F, Ren L, Xiong Z, Wu Z, Dong J, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol*. 2011;49:3463-9.
- [18] Nasu A, Marusawa H, Ueda Y, Nishijima N, Takahashi K, Osaki Y, et al. Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. *PLoS ONE*. 2011;6:e24907.
- [19] Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci USA*. 2008;105:13081-6.
- [20] De Grassi A, Segala C, Iannelli F, Volorio S, Bertario L, Radice P, et al. Ultradeep Sequencing of a Human Ultraconserved Region Reveals Somatic and Constitutional Genomic Instability. *PLoS Biol*. 2010;8:e1000275.
- [21] Zagordi O, Klein R, Däumer M, Beerwinkler N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Research*. 2010;38:7400-9.
- [22] Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*. 2007;17:1195-201.

- [23] Carlson CA, Kas A, Kirkwood R, Hays LE, Preston BD, Salipante SJ, et al. Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nat Methods*. 2012;9:78-80.
- [24] He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, et al. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*. 2010;464:610-4.
- [25] Ameer A, Stewart JB, Freyer C, Hagström E, Ingman M, Larsson N-G, et al. Ultra-Deep Sequencing of Mouse Mitochondrial DNA: Mutational Patterns and Their Origins. *PLoS Genet*. 2011;7:e1002028.
- [26] Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng*. 2003;96:317-23.
- [27] Meyerhans A, Vartanian JP, Wain-Hobson S. DNA recombination during PCR. *Nucleic Acids Research*. 1990;18:1687-91.
- [28] Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods*. 2008;5:1005-10.
- [29] Salk J, Fox E, Loeb L. Mutational heterogeneity in human cancers: origin and consequences. *Annual Review of Pathology*. 2009;5:51-75.
- [30] Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*. 2009;6:291-5.
- [31] Vandenbroucke I, Van Marck H, Verhasselt P, Thys K, Mostmans W, Dumont S, et al. Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. *BioTechniques*. 2011;51:167-77.
- [32] Flaherty P, Natsoulis G, Muralidharan O, Winters M, Buenrostro J, Bell J, et al. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Research*. 2012;40:e2-e.
- [33] Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, et al. A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res*. 2010;20:273-80.
- [34] Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Research*. 2004;32:e135.
- [35] McCloskey ML, Stöger R, Hansen RS, Laird CD. Encoding PCR products with batch-stamps and barcodes. *Biochem Genet*. 2007;45:761-7.

- [36] Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA*. 2011;108:9530-5.
- [37] Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci USA*. 2011;108:20166-71.
- [38] Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*. 2011;9:72-4.
- [39] Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*. 2011;39:e81-e.
- [40] Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci USA*. 2012;109:1347-52.
- [41] Fu GK, Hu J, Wang P-H, Fodor SPA. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci USA*. 2011;108:9026-31.
- [42] Cervantes RB, Stringer JR, Shao C, Tischfield JA, Stambrook PJ. Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proc Natl Acad Sci USA*. 2002;99:3586-90.
- [43] Lindahl T, Wood RD. Quality control by DNA repair. *Science*. 1999;286:1897-1905.
- [44] Kunkel, TA. Mutational specificity of depurination. *Proc Natl Acad Sci USA*. 1984;81:1494-98.
- [45] Shibutani S, Takeshita M, Grollman AP. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature*. 1991;349:431-4.
- [46] Stiller M, Green RE, Ronan M, Simons JF, Du L, He W., et al. Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc Natl Acad Sci USA*. 2006;103:13578-84.
- [47] Ehrich M, Deciu C, Zwiefelhofer T, Tynan JA, Cagasan L, Tim R, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. *Am J Obsbet Gynecol*. 2011;204:205e1-11.
- [48] Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch M, D'Ascenzo M, et al. Whole exome capture in solution with 3 Gbp of data. *Genome Biol*. 2010;11:R62:1-8.

[49] Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 2010; 38:159e1---8.

[50] Kaur M, Makrigiorgos GM. Novel amplification of DNA in a hairpin structure: towards a radical elimination of PCR errors from amplified DNA. *Nucleic Acids Res.* 2003; 31:26e1---7.

[51] Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., and Milos, P.M. (2009). Direct RNA sequencing. *Nature* 461, 814–818.

## CLAIMS

What is claimed is

1. A single molecule identifier adaptor molecule for use in sequencing a double-stranded target nucleic acid molecule comprising  
  
a double-stranded single molecule identifier (SMI) sequence, the double-stranded SMI sequence comprising a double-stranded degenerate or semi-degenerate nucleic acid sequence; and  
  
an SMI ligation adaptor that allows the SMI adaptor molecule to be ligated to the double-stranded target nucleic acid sequence.
2. The single molecule identifier adaptor molecule of claim 1, wherein the double-stranded target nucleic acid molecule is a double-stranded DNA or RNA molecule.
3. The single molecule identifier adaptor molecule of claim 1, further comprising at least two PCR primer binding sites, or at least two sequencing primer binding sites, or both.
4. The single molecule identifier adaptor molecule of claim 1, further comprising a double-stranded fixed reference sequence
5. The single molecule identifier adaptor molecule of claim 1, wherein the double-stranded degenerate or semi-degenerate nucleic acid sequence comprises a first nucleotide n-mer sequence and a second n-mer sequence that is complementary to the first nucleotide n-mer sequence.

6. The single molecule identifier adaptor molecule of claim 5, wherein the first n-mer sequence comprises a nucleotide sequence that is between approximately 4 and 20 nucleotides in length.
7. The single molecule identifier adaptor molecule of claim 5, wherein the first nucleotide n-mer sequence is a degenerate sequence.
8. The single molecule identifier adaptor molecule of claim 1, wherein the double-stranded degenerate or semi-degenerate DNA sequence comprises a randomly fragmented double stranded nucleic acid derived from an alternative source.
9. The single molecule identifier adaptor molecule of claim 1, wherein the SMI ligation adaptor is selected from a T-overhang, an A-overhang, a CG overhang, a blunt end, or any other ligatable nucleic acid sequence.
10. The single molecule identifier adaptor molecule of claim 1, wherein the SMI adaptor molecule is Y-shaped, U-shaped, or a combination thereof.
11. A method of obtaining the sequence of a double-stranded target nucleic acid comprising
  - ligating a double-stranded target nucleic acid molecule to at least one SMI adaptor molecule to form a double-stranded SMI-target nucleic acid complex;
  - amplifying the double-stranded SMI-target nucleic acid complex, resulting in a set of amplified SMI-target nucleic acid products; and
  - sequencing the amplified SMI-target nucleic acid products.

12. The single molecule identifier adaptor molecule of claim 11, wherein the double-stranded target nucleic acid molecule is a double-stranded DNA or RNA molecule.

13. The method of claim 11, further comprising generating an error-corrected double-stranded consensus sequence by (i) grouping the sequenced SMI-target nucleic acid products into families of paired target nucleic acid strands based on a common set of SMI sequences; and (ii) removing paired target nucleic acid strands having one or more nucleotide positions where the paired target nucleic acid strands disagree, or alternatively removing nucleotide positions from nucleic acid strands where the paired strands disagree at that specific position.

14. The method of claim 11, wherein the double-stranded target nucleic acid molecule is a sheared double-stranded DNA or RNA fragment.

15. The method of claim 14, wherein the sheared double-stranded nucleic acid fragment further comprises a double-stranded target nucleic acid sequence ligation adaptor.

16. The method of claim 15, wherein the double-stranded target nucleic acid sequence ligation adaptor is selected from a T-overhang, an A-overhang, a CG overhang, a blunt end, or any ligatable nucleic acid sequence

17. The method of claim 16, wherein each end of the double-stranded target nucleic acid molecule is ligated to an SMI adaptor molecule.

18. The method of claim 17, wherein each SMI adaptor molecules comprises

a double-stranded single molecule identifier (SMI) sequence, the double-stranded SMI sequence comprising a double-stranded degenerate or semi-degenerate nucleic acid sequence; and

an SMI ligation adaptor that allows the SMI adaptor molecule to be ligated to the double-stranded target nucleic acid sequence.

19. The method of claim 18, further comprising at least two PCR primer binding sites, at least two sequencing primer binding sites, a double-stranded fixed reference sequence or a combination thereof.

20. The single molecule identifier adaptor molecule of claim 18, wherein the double-stranded degenerate or semi-degenerate nucleic acid sequence comprises a first nucleotide n-mer sequence and a second n-mer sequence that is complementary to the first nucleotide n-mer sequence.

**ABSTRACT**

Next Generation DNA sequencing promises to revolutionize clinical medicine and basic research. However, while this technology has the capacity to generate hundreds of billions of nucleotides of DNA sequence in a single experiment, the error rate of approximately 1% results in hundreds of millions of sequencing mistakes. These scattered errors can be tolerated in some applications but become extremely problematic when “deep sequencing” genetically heterogeneous mixtures, such as tumors or mixed microbial populations. To overcome limitations in sequencing accuracy, a method Duplex Consensus Sequencing (DCS) is provided. This approach greatly reduces errors by independently tagging and sequencing each of the two strands of a DNA duplex. As the two strands are complementary, true mutations are found at the same position in both strands. In contrast, PCR or sequencing errors will result in errors in only one strand. This method uniquely capitalizes on the redundant information stored in double-stranded DNA, thus overcoming technical limitations of prior methods utilizing data from only one of the two strands.

Figure 1

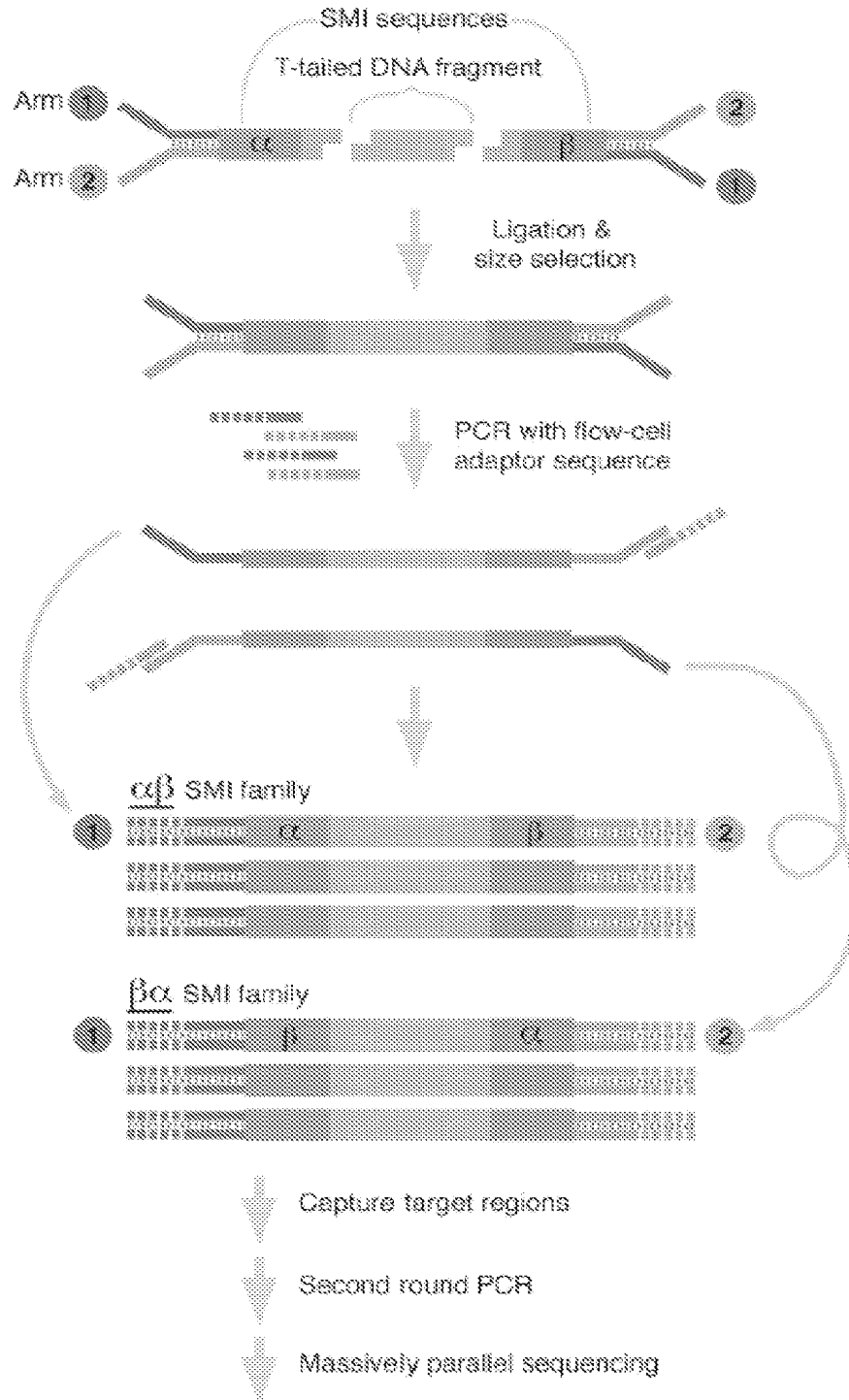


Figure 2

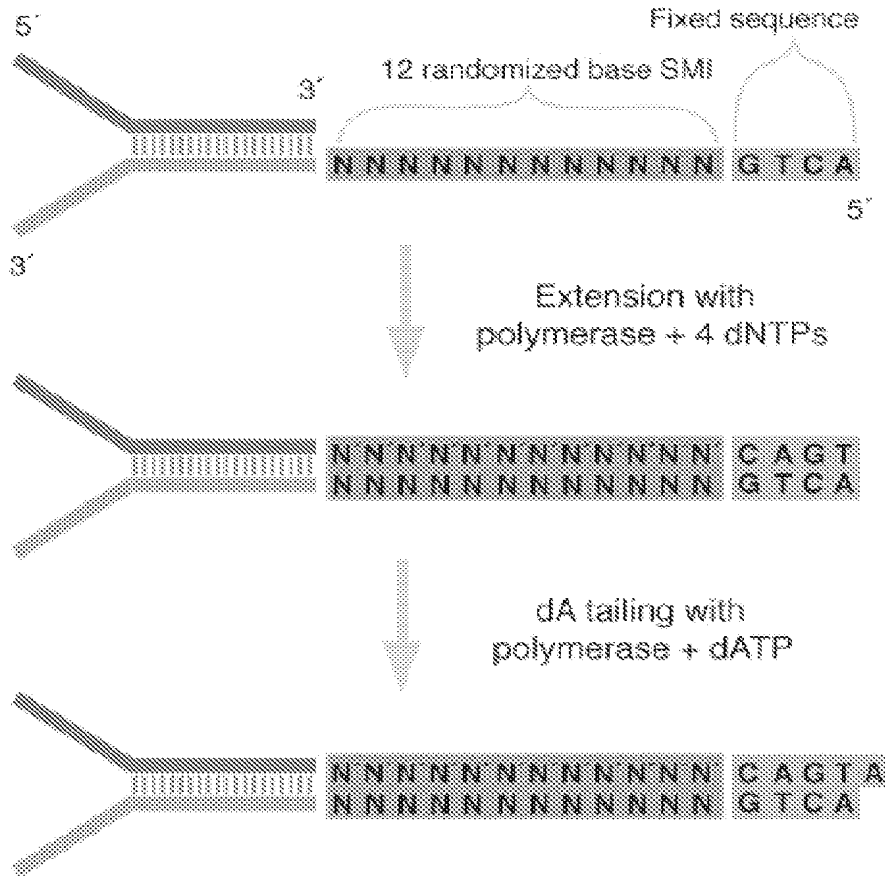
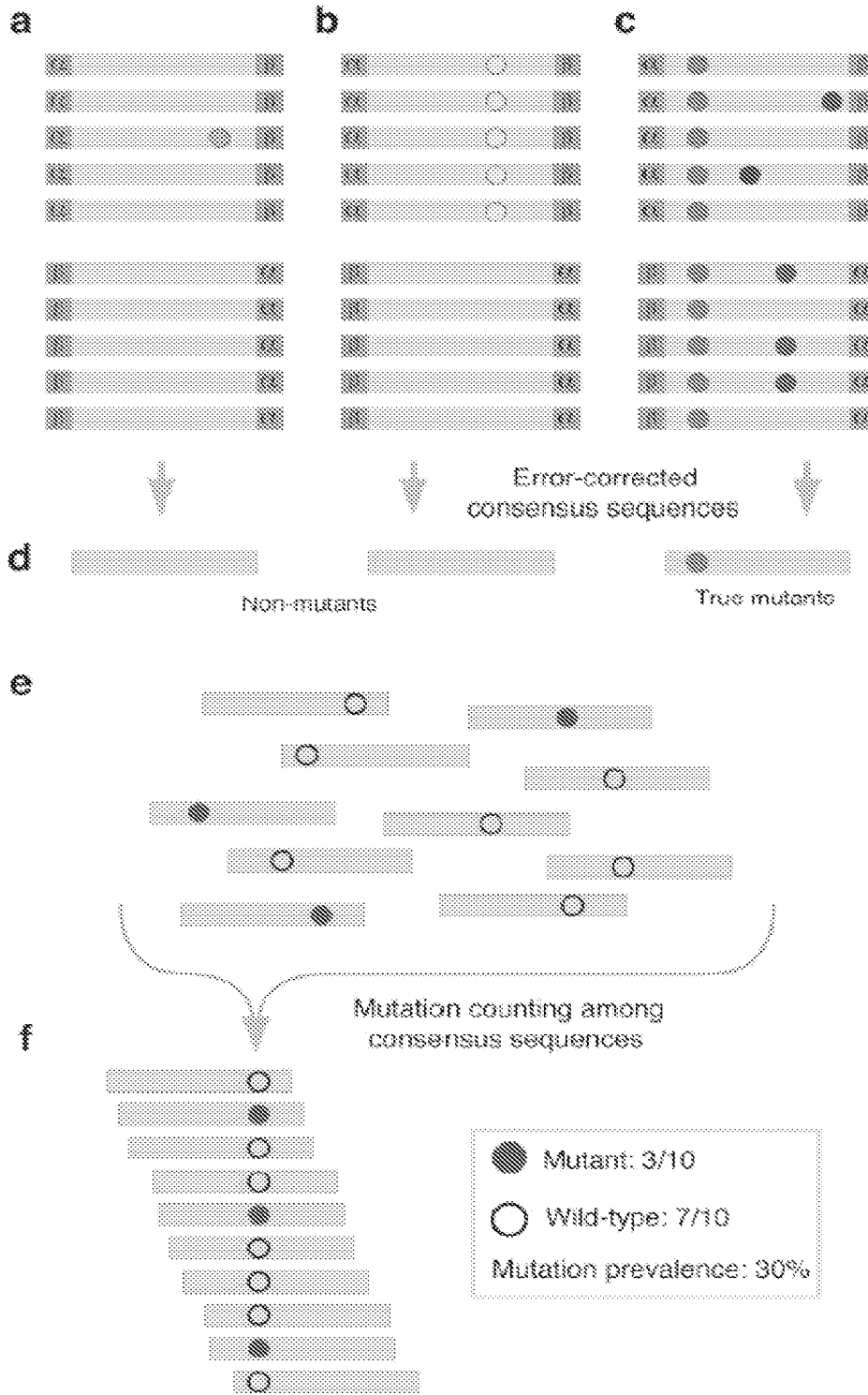


Figure 3



## Electronic Acknowledgement Receipt

<b>EFS ID:</b>	12352131
<b>Application Number:</b>	61613413
<b>International Application Number:</b>	
<b>Confirmation Number:</b>	1819
<b>Title of Invention:</b>	METHODS OF LOWERING THE ERROR RATE OF MASSIVELY PARALLEL DNA SEQUENCING USING DUPLEX CONSENSUS SEQUENCING
<b>First Named Inventor/Applicant Name:</b>	Michael Schmitt
<b>Customer Number:</b>	94991
<b>Filer:</b>	Lara J. Dueppen/Amy Shields
<b>Filer Authorized By:</b>	Lara J. Dueppen
<b>Attorney Docket Number:</b>	72227.8043.US00
<b>Receipt Date:</b>	20-MAR-2012
<b>Filing Date:</b>	
<b>Time Stamp:</b>	19:47:31
<b>Application Type:</b>	Provisional

### Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	\$ 125
RAM confirmation Number	6405
Deposit Account	502586
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:  
Charge any Additional Fees required under 37 C.F.R. Section 1.17 (Patent application and reexamination processing fees)

<b>File Listing:</b>					
<b>Document Number</b>	<b>Document Description</b>	<b>File Name</b>	<b>File Size(Bytes)/ Message Digest</b>	<b>Multi Part /.zip</b>	<b>Pages (if appl.)</b>
1	Provisional Cover Sheet (SB16)	2012-03-20_PRO_CoverSheet_722278043US.PDF	1000852 cf0eccc/c/e1966f336549f288e2/08c6/3648de	no	3
<b>Warnings:</b>					
<b>Information:</b>					
2		2012-03-20_Specification_722278043US.pdf	251570 a4f254765104b431d4c5f1597ae191d2d14053db	yes	47
	<b>Multipart Description/PDF files in .zip description</b>				
	<b>Document Description</b>		<b>Start</b>	<b>End</b>	
	Specification		1	42	
	Claims		43	46	
	Abstract		47	47	
<b>Warnings:</b>					
<b>Information:</b>					
3	Drawings-other than black and white line drawings	2012-03-20_Drawings_722278043US.pdf	263208 3298f0141ef9f9832fe9d446a0dc16206ffcb20a	no	3
<b>Warnings:</b>					
<b>Information:</b>					
4	Fee Worksheet (SB06)	fee-info.pdf	30071 9991a380ac48b8b91d3a338866ff3ca6fc0ffc41	no	2
<b>Warnings:</b>					
<b>Information:</b>					
<b>Total Files Size (in bytes):</b>			1545701		

**This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.**

**New Applications Under 35 U.S.C. 111**

**If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.**

**National Stage of an International Application under 35 U.S.C. 371**

**If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.**

**New International Application Filed with the USPTO as a Receiving Office**

**If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.**

## SCORE Placeholder Sheet for IFW Content

Application Number: 61613413

Document Date: 3/20/2012

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

- Drawings

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

To access the documents in the SCORE database, refer to instructions developed by SIRA.

At the time of document entry (noted above):

- Examiners may access SCORE content via the eDAN interface using the Supplemental Content tab.
- Other USPTO employees can bookmark the current SCORE URL (<http://es/ScoreAccessWeb/>).
- External customers may access SCORE content via the Public and Private PAIR interfaces using the Supplemental Content tab.

## **DOCUMENT MADE AVAILABLE UNDER THE PATENT COOPERATION TREATY (PCT)**

International application number:	<b>PCT/US2013/032665</b>
International filing date:	<b>15 March 2013 (15.03.2013)</b>
Document type:	<b>Certified copy of priority document</b>
Document details:	Country/Office: <b>US</b>
	Number: <b>61/613,413</b>
	Filing date: <b>20 March 2012 (20.03.2012)</b>
Date of receipt at the International Bureau:	<b>01 July 2014 (01.07.2014)</b>

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a),(b) or (b-bis)