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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 September 2013 (26.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/142389 A1

(51) International Patent Classification:
C04B 20/04 (2006.01)

(21) International Application Number:
PCT/US2013/032665

(22) International Filing Date:
15 March 2013 (15.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/613,413	20 March 2012 (20.03.2012)	US
61/625,623	17 April 2012 (17.04.2012)	US
61/625,319	17 April 2012 (17.04.2012)	US

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

(72) Inventors: **SCHMITT, Michael**; Seattle, WA 98105 (US). **SALK, Jesse**; Seattle, WA 98105 (US). **LOEB, Lawrence, A.**; Bellevue, WA (US).

(74) Agent: **DUEPPEN, Lara, J.**; Perkins Coie LLP, P.O. Box 1208, Seattle, WA 98111-1208 (US).

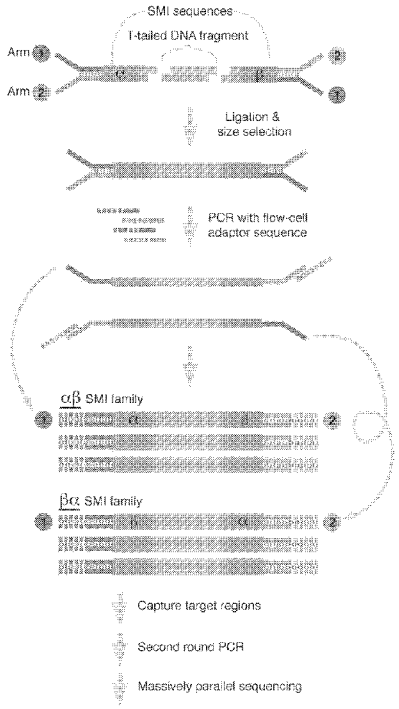
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE,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

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

[Continued on next page]

(54) Title: METHODS OF LOWERING THE ERROR RATE OF MASSIVELY PARALLEL DNA SEQUENCING USING DUPLEX CONSENSUS SEQUENCING

Figure 1



(57) 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.

WO 2013/142389 A1

Published:

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

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

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

PRIORITY CLAIM

[0001] This application claims priority to U.S. Provisional Patent Application No. 61/613,413, filed March 20, 2012; U.S. Provisional Patent Application No. 61/625,623, filed April 17, 2012; and U.S. Provisional Patent Application No. 61/625,319, filed April 17, 2012; the subject matter of all of which are hereby incorporated by reference as if fully set forth herein.

STATEMENT OF GOVERNMENT INTEREST

[0002] 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

[0003] 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].

[0004] This concept of "deep sequencing" has been implemented in a variety 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 mosasism 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.

[0005] 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 stoichastic and non-stoichastic 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].

[0006] 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.

[0007] 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].

[0008] 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

[0009] In one embodiment, a single molecule identifier (SMI) adaptor molecule for use in sequencing a double-stranded target nucleic acid molecule is provided. Said

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.