Detection of ultra-rare mutations by next-generation sequencing

Michael W. Schmitt[a], Scott R. Kennedy[a], Jesse J. Salk[a], Edward J. Fox[a], Joseph B. Hiatt[b], and Lawrence A. Loeb[a,c,1]

Departments of [a]Pathology, [b]Genome Sciences, and [c]Biochemistry, University of Washington School of Medicine, Seattle, WA 98195

Edited* by Mary-Claire King, University of Washington, Seattle, WA, and approved July 3, 2012 (received for review June 6, 2012)

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 ∼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, we have developed a method termed Duplex Sequencing. 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 result in mutations in only one strand and can thus be discounted as technical error. We determine that Duplex Sequencing has a theoretical background error rate of less than one artifactual mutation per billion nucleotides sequenced. In addition, we establish that detection of mutations present in only one of the two strands of duplex DNA can be used to identify sites of DNA damage. We apply the method to directly assess the frequency and pattern of random mutations in mitochondrial DNA from human cells.

cancer | diagnostics | subclone | quasispecies | biomarker

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 a small fraction of the time and cost of traditional Sanger methods (1). Unlike conventional techniques, which simply report the average genotype of an aggregate collection of molecules, next-generation sequencing technologies digitally tabulate the sequence of many individual DNA fragments, thus offering the unique ability to detect minor variants within heterogeneous mixtures. This concept of "deep sequencing" has been implemented in a variety of fields including metagenomics (2), paleogenomics (3), forensics (4), and human genetics (5) to disentangle subpopulations in complex biological samples. Clinical applications are rapidly being developed, such as prenatal screening for fetal aneuploidy (6), early detection of cancer (7), and monitoring its response to therapy (8) with nucleic acid-based serum biomarkers.

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 (9). PCR amplification of heterogeneous mixtures can result in population skewing due to differential amplification (10, 11), and polymerase mistakes generate point mutations resulting from base misincorporations and rearrangements due to template switching (10, 12). Combined with the additional errors that arise during cluster amplification, cycle sequencing, and image analysis, ∼1% of bases are incorrectly identified, depending on the specific platform and sequence context (1). This background level of artifactual heterogeneity establishes a limit below which the presence of true rare variants is obscured (9).

A variety of improvements at the level of biochemistry (13, 14) and data processing (14–19) have been developed to improve sequencing accuracy. In addition, techniques whereby PCR duplicates arising from individual DNA fragments can be resolved on the basis of unique random shear points (20) or via exogenous tagging (21, 22) before amplification (23–28) have recently been reported. Because all amplicons derived from a particular starting molecule can be explicitly identified, 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 (25, 26, 28) and RNA templates (24, 25, 27, 29) and to correct base errors arising during PCR or sequencing (20, 23, 24, 26). For example, Kinde et al. (23) reported a reduction in error frequency of ∼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 approach allowed for an observed mutation frequency of ∼0.001% mutations/bp in normal human genomic DNA. 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^{-8} to 10^{-11} (30, 31). Thus, the majority of mutations seen in normal human genomic DNA by this method potentially still represent technical artifacts.

Prevailing next-generation sequencing platforms generate sequence data from single-stranded fragments of DNA. As a consequence, artifactual mutations introduced during the initial round of PCR amplification are undetectable as errors—even with tagging techniques—if the base change is propagated to all subsequent PCR duplicates. Multiple 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 (32), and additional DNA damage is generated ex vivo during tissue processing and DNA extraction (33).

Limitations inherent to sequencing of single-stranded DNA can be overcome, however, as DNA naturally exists as a double-stranded entity, with one molecule reciprocally encoding the sequence information of its partner. Thus, it should be feasible to identify and correct nearly all forms of sequencing errors by comparing the sequence of individual tagged amplicons derived from one half of a double-stranded complex with those of the other half of the same molecule. Herein, we present an approach for tag-based error correction, termed Duplex Sequencing, which capitalizes on the redundant information stored in complexed double-stranded DNA. Our method has a theoretical background error rate of less than one artifactual error per 10^9 nucleotides sequenced and thus allows rare variants in heterogeneous populations to be detected with unprecedented sensitivity.

Author contributions: M.W.S., S.R.K., J.J.S., and L.A.L. designed research; M.W.S., S.R.K., and E.J.F. performed research; M.W.S., S.R.K., and J.B.H. contributed new reagents/analytic tools; M.W.S., S.R.K., J.J.S., and L.A.L. analyzed data; and M.W.S., S.R.K., J.J.S., and L.A.L. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

See Commentary on page 14289.

[1]To whom correspondence should be addressed. E-mail: laloeb@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1208715109/-/DCSupplemental.

IDT Exhibit 1031.0001

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

SEE COMMENTARY

Results

To improve the sensitivity of variant detection by next-generation DNA sequencing, we designed an alternative approach to library preparation and analysis that we term Duplex Sequencing. The method entails tagging both strands of duplex DNA with a random, yet complementary double-stranded nucleotide sequence, which we refer to as a Duplex Tag. Double-stranded tag sequences are incorporated into standard Illumina sequencing adapters by first introducing a single-stranded randomized nucleotide sequence into one adapter strand and then extending the opposite strand with a DNA polymerase to yield a complementary, double-stranded tag (Fig. 1*A*). Following ligation of tagged adapters to sheared DNA, the individually labeled strands are PCR amplified from asymmetric primer sites on the adapter tails (Fig. 1*B*) and subjected to paired-end sequencing. Every PCR duplicate that arises from a single strand of DNA will carry the original strand's tag sequence. Owing to the complementary nature of the Duplex Tags on the two strands, each strand in a DNA duplex pair generates a distinct, yet related, population of PCR duplicates. Comparing the sequence obtained from each of the two strands in a duplex facilitates differentiation of sequencing errors from true mutations: when an apparent mutation is, in fact, 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.

During the PCR amplification step after tagging, many duplicate "families" of molecules are created, each of which arose from a single strand of an individual DNA molecule. After sequencing, members of each PCR family are identified and grouped by virtue of sharing an identical tag sequence (Fig. 1*C*). The sequences of uniquely tagged PCR duplicates are then compared to create a PCR consensus sequence. Only PCR families consisting of at least three duplicates and yielding the same sequence in at least 90% of the members at a given position are used to create the consensus sequence. This step filters out random errors introduced during sequencing or PCR to yield a set of sequences, each of which derives from an individual molecule of single-stranded DNA. We refer to these as single strand consensus sequences (SSCSs).

Next, sequences belonging to the two complementary strands of each DNA duplex are identified by searching for complementary tag sequences among SSCS reads. Specifically, a 24-nucleotide tag sequence consists of two 12-nucleotide sequences at each end of the molecule that can be designated α and β. For a tag of form αβ in read 1, the opposite strand's tag will be of form βα in read 2. Following partnering of the two strands, the sequences of the strands are compared. A sequence base at a given position is kept only if the read data from each of the two strands matches perfectly. A detailed illustration of the approach is provided in *SI Materials and Methods*. Comparing the sequences obtained from both strands eliminates errors introduced during the first round of PCR where an artifactual mutation may be propagated to all PCR duplicates of one strand and would not be removed by SSCS filtering alone. We refer to the resulting high-confidence sequences of individual DNA duplex molecules as duplex consensus sequences (DCSs).

Duplex Sequencing of M13 DNA. To establish the sensitivity of Duplex Sequencing, we first applied the method to M13mp2

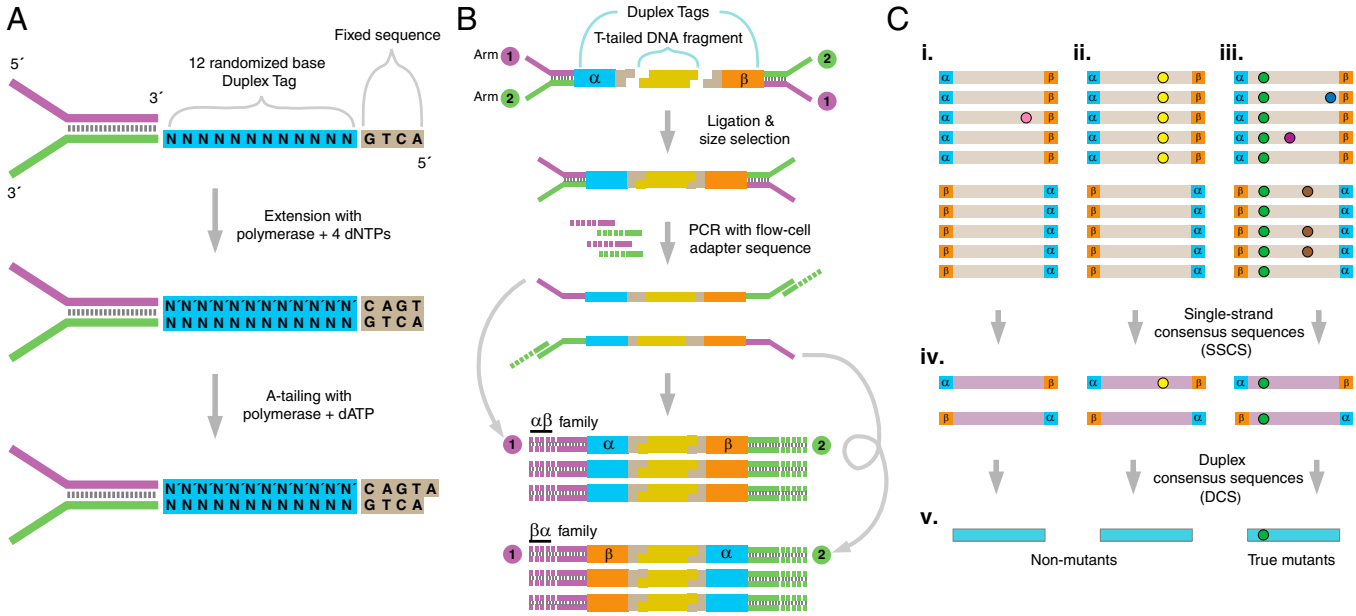

Fig. 1. Overview of Duplex Sequencing. (*A*) Adapter synthesis. A double-stranded, randomized Duplex Tag sequence is appended to a sequencing adapter by copying a degenerate sequence in one strand of the adapter with DNA polymerase. Complete adapter A-tailing is ensured by extended incubation with polymerase and dATP. (*B*) Duplex Sequencing workflow. Sheared, T-tailed double-stranded DNA is ligated to A-tailed adapters. Because every adapter contains a Duplex Tag on each end, every DNA fragment becomes labeled with two distinct tag sequences (arbitrarily designated α and β in the single fragment shown). PCR amplification with primers containing Illumina flow-cell–compatible tails is carried out to generate families of PCR duplicates. Two types of PCR products are produced from each DNA fragment. Those derived from one strand will have the α tag sequence adjacent to flow cell sequence 1 and the β tag sequence adjacent to flow cell sequence 2. PCR products originating from the complementary strand are labeled reciprocally. (*C*) Error correction. (*i–iii*) Sequence reads sharing a unique set of tags are grouped into paired families with members having strand identifiers in either the αβ or βα orientation. Each family pair reflects the amplification of one double-stranded DNA fragment. (*i*) Mutations (colored spots) present in only one or a few family members represent sequencing mistakes or PCR-introduced errors occurring late in amplification. (*ii*) Mutations occurring in many or all members of one family in a pair arise from PCR errors during the first round of amplification such as might occur when copying across sites of mutagenic DNA damage. (*iii*) True mutations (green) present on both strands of a DNA fragment appear in all members of a family pair. Whereas artifactual mutations may co-occur in a family pair with a true mutation, all except those arising during the first round of PCR amplification can be independently identified and discounted when producing (*iv*) an error-corrected single-strand consensus sequence (SSCS). The sequences obtained from each of the two strands of an individual DNA duplex can then be compared to obtain (*v*) the duplex consensus sequence (DCS), which eliminates remaining errors that occurred during the first round of PCR.

GENETICS

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1031.0002

DNA, which is a substrate that has been used extensively in sensitive genetic mutation assays and has a well-established base substitution frequency of 3.0×10^{-6} (34). M13mp2 DNA was sheared and ligated to Duplex Sequencing adapters and subjected to deep sequencing on an Illumina HiSEq 2000 (Fig. 2*A*). Analysis of the data by standard methods (i.e., without consideration of the double-stranded tag sequences and with quality filtering for a Phred score of 30) resulted in an error frequency of 3.8×10^{-3}, more than 1,000-fold higher than the true mutation frequency of M13mp2 DNA. Thus, >99.9% of the apparent mutations identified by standard sequencing are erroneous.

We generated SSCSs by using the unique tag affixed to each molecule to create a consensus of all PCR products that came from an individual molecule of single-stranded DNA. This resulted in a mutation frequency of 3.4×10^{-5}, suggesting that ~99% of sequencing errors are corrected in SSCS reads. However, this mutation frequency is >10-fold higher than the reference value of 3.0×10^{-6}, indicating that ~90% of the mutations identified by SSCSs are still artifacts.

Next, we further corrected errors by using the complementary tags to compare the DNA sequence arising from the two strands of each single molecule of duplex DNA to create DCSs. This approach resulted in a mutation frequency of 2.5×10^{-6}, nearly identical to the frequency of 3.0×10^{-6} determined by well-established genetic methods (34). The number of nucleotides of DNA sequence obtained by a standard sequencing approach, and after SSCS and DCS analysis, may be found in Table S1.

DNA Damage Alters SSCS Mutation Spectrum. We next examined the spectrum of mutations identified by both SSCS and DCS analysis relative to literature reference values (34) for the M13mp2 substrate (Fig. 2*B*). SSCS analysis revealed a large excess of G→A/C→T and G→T/C→A mutations relative to reference ($P < 10^{-6}$, two-sample *t* test). In contrast, DCS analysis was in excellent agreement with the literature values with the exception of a *decrease* relative to reference of these same mutational events: G→A/C→T and G→T/C→A ($P < 0.01$). To probe the potential cause of these spectrum deviations, the SSCS data were filtered to consist of forward-mapping reads from read 1 (i.e., direct sequencing of the reference strand) and the reverse complement of reverse-mapping reads from read 1 (i.e., direct sequencing of the antireference strand.) True double-stranded mutations should result in an equal balance of complementary mutations observed on the reference and antireference strand. However, SSCS analysis revealed a large number of single-stranded G→T mutations, with a much smaller number of C→A mutations (Fig. 2*C*). A similar bias was seen with a large excess of C→T mutations relative to G→A mutations.

Base-specific mutagenic DNA damage is a likely explanation of these imbalances. Excess G→T mutations are consistent with the oxidative product 8-oxo-guanine (8-oxo-G) causing first round PCR errors and artifactual G→T mutations. DNA polymerases, including those commonly used in PCR, have a strong tendency to insert adenine opposite 8-oxo-G (35, 36), and misinsertion of A opposite 8-oxo-G would result in erroneous scoring of a G→T mutation. Likewise, the excess C→T mutations are consistent with spontaneous deamination of cytosine to uracil (37), a particularly common DNA damage event that results in insertion during PCR of adenine opposite uracil and erroneous scoring of a C→T mutation.

To determine whether the excess G→T mutations seen in SSCSs might reflect oxidative DNA damage at guanine nucleotides, before sequencing library preparation we incubated M13mp2 DNA with the free radical generator hydrogen peroxide in the presence of iron, a protocol that induces DNA damage (38). This treatment resulted in a substantial further increase in G→T mutations by SSCS analysis (Fig. 3*A*), consistent with PCR errors at sites of DNA damage as the likely mechanism of this biased mutation spectrum. In contrast, induction of oxidative damage did not alter the mutation spectrum seen with DCS analysis (Fig. 3*B*),

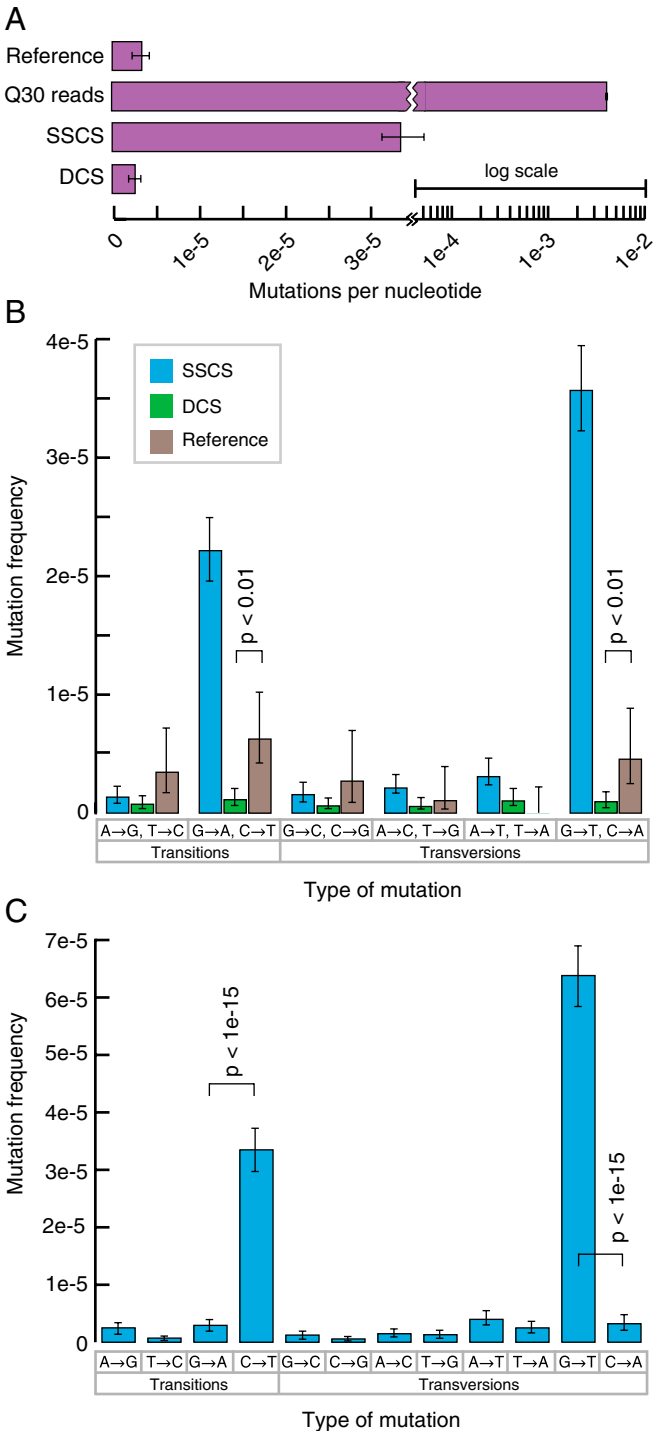

Fig. 2. Duplex Sequencing of M13mp2 DNA. (*A*) Average mutation frequency of M13mp2 DNA as measured by a standard sequencing approach, SSCS, and DCS. Reference value of 3.0×10^{-6} is from ref. 34. Note that the axis is plotted on a split-log scale. (*B*) Single-strand consensus sequences (SSCSs) reveal a large excess of G→A/C→T and G→T/C→A mutations, whereas duplex consensus sequences (DCSs) yield a balanced spectrum. Mutation frequencies are grouped into reciprocal mispairs, as DCS analysis only scores mutations present in both strands of duplex DNA. All significant ($P < 0.05$) differences between DCS analysis and the literature reference values are noted. (*C*) Complementary types of mutations should occur at approximately equal frequencies within a DNA fragment population derived from duplex molecules. However, SSCS analysis yields a 15-fold excess of G→T mutations relative to C→A mutations and an 11-fold excess of C→T mutations relative to G→A mutations. All significant ($P < 0.05$) differences between paired reciprocal mutation frequencies are noted.

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1031.0003

SEE COMMENTARY

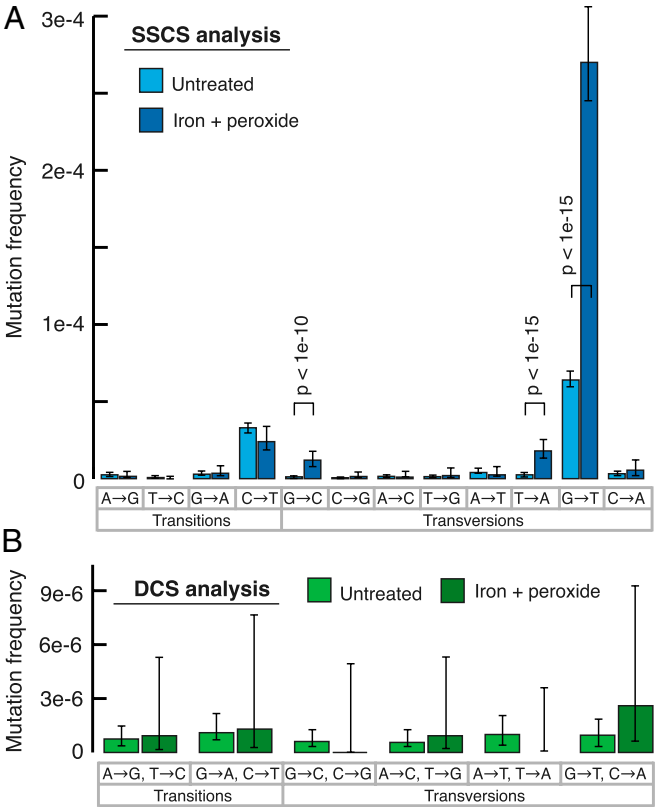

Fig. 3. Effect of DNA damage on mutation spectrum. DNA damage was induced by incubating purified M13mp2 DNA with hydrogen peroxide and $FeSO_4$. (*A*) SSCS analysis reveals a further elevation from baseline of G→T mutations, indicating these events to be the artifactual consequence of nucleotide oxidation. All significant ($P < 0.05$) changes from baseline mutation frequencies are noted. (*B*) Induced DNA damage had no effect on the overall frequency or spectrum of DCS mutations.

indicating that duplex consensus sequences are not similarly susceptible to DNA damage artifacts.

Furthermore, relative to the literature reference values, DCS analysis results in a lower frequency of G→T/C→A and C→T/G→A mutations (Fig. 2*B*), which are the same mutations elevated in SSCS analysis as a probable result of DNA damage. Notably, the M13mp2 LacZ assay, from which reference values have been derived, is dependent upon bacterial replication of a single molecule of M13mp2 DNA. Thus, the presence of oxidative damage within this substrate could cause an analogous first-round replication error by *Escherichia coli*, converting a single-stranded damage event into a fixed, double-stranded mutation during replication. The slight reduction in the frequency of these two types of mutations measured by DCS analysis may, therefore, reflect the absence of damage-induced errors that are scored by the in vivo LacZ assay.

Mutant Recovery. To further validate the capability of DCS analysis to detect rare mutations, we constructed a series of M13mp2 variants containing specific single base substitutions and mixed the variants together at known ratios. The final mixture was then sequenced with Duplex Sequencing adapters. With conventional analysis of the sequencing data (i.e., without consideration of the tag sequences and filtering for a read quality score of 30), variants present at a level of <1/100 could not be accurately identified because artifactual mutations occurring at a background frequency of about 1/100 obscured the presence of less abundant true mutations (Fig. S1). In contrast, when the data are analyzed as duplex consensus sequences with ~20,000-fold final depth, accurate recovery of mutant sequences was possible down to the lowest tested level of one mutant molecule per 10,000 wild-type molecules.

Duplex Sequencing of Human Mitochondrial DNA. Having established the methodology for Duplex Sequencing with M13mp2 DNA, which is a substrate for which the mutation frequency and spectrum are fairly well established, we next wished to apply the approach to a human DNA sample. Thus, we isolated mitochondrial DNA from human brain tissue and sequenced the DNA after ligation of Duplex Sequencing adapters. A standard sequencing approach with quality filtering for a Phred score of 30 resulted in a mutation frequency of 2.7×10^{-3}, and SSCS analysis yielded a mutation frequency of 1.5×10^{-4}. In contrast, DCS analysis revealed a much lower overall mutation frequency of 3.5×10^{-5} (Fig. 4*A*). The frequency of mutations in mitochondrial DNA has previously been difficult to measure directly due in part to sources of error in existing assays that can result in either overestimation or underestimation of the true value. An additional confounder has been that most approaches are limited to interrogation of mutations within a small fraction of the genome (39). The method of single-molecule PCR, which has been proposed as an accurate method of measuring mitochondrial mutation frequency (39) and is considered resistant to damage-induced background errors (40), has resulted in a reported mitochondrial mutation frequency in human colonic mucosa of $5.9 \times 10^{-5} \pm 3.2 \times 10^{-5}$ (39), which is in excellent agreement with our result. Likewise, mitochondrial DNA sequence divergence rates in human pedigrees are consistent with a mitochondrial mutation frequency of $3–5 \times 10^{-5}$ (41, 42).

When the distribution of mutations throughout the mitochondrial genome is considered, the quality filtered reads (analyzed without consideration of the tags) have many artifactual errors, such that identification of mutational hotspots is difficult or impossible (Fig. 4*B*). DCS analysis removed these artifacts (Fig. 4*C*) and revealed striking hypermutability of the region of replication initiation (D loop), which is consistent with prior estimates of mutational patterns in mitochondrial DNA based upon sequence variation at this region within the population (43).

SSCS analysis produced a strong mutational bias, with a 130-fold excess of G→T relative to C→A mutations (Fig. 4*D*), consistent with oxidative damage of the DNA leading to first-round PCR mutations as a significant source of background error. A high level of oxidative damage is expected in mitochondrial DNA, due to extensive exposure of mitochondria to free radical species generated as a byproduct of metabolism (44). DCS analysis (Fig. 4*E*) removed the mutational bias and revealed that transition mutations are the predominant replication errors in mitochondrial DNA. The DCS mutation spectrum is in accord with prior estimates of deamination events (45) and T-dGTP mispairing by the mitochondrial DNA polymerase (46) as primary mutational forces in mitochondrial DNA. Furthermore, the mutation spectrum of our mitochondrial data are consistent with previous reports of heteroplasmic mutations in human brain showing an increased load of A→G/T→C and G→A/C→T transitions, relative to transversions (47, 48). A similar spectral bias has also been reported in mice (45, 49) and in population studies of *Drosophila melanogaster* (50).

Discussion

The accuracy of standard approaches to next-generation sequencing is constrained by a general reliance on analysis of single-stranded DNA, which makes certain technical sources of single-stranded errors fundamentally limiting. The complementary strands of native duplex DNA harbor redundant sequence information and here we have demonstrated an approach for error correction, termed Duplex Sequencing, which capitalizes on this biochemical redundancy to greatly lower the error rate of sequencing.

The most sensitive approach previously reported for improving accuracy of next-generation sequencing involves use of a random tag sequence in a PCR primer (23). In this technique, PCR duplicates are generated from a single strand of DNA, and the

GENETICS

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1031.0004

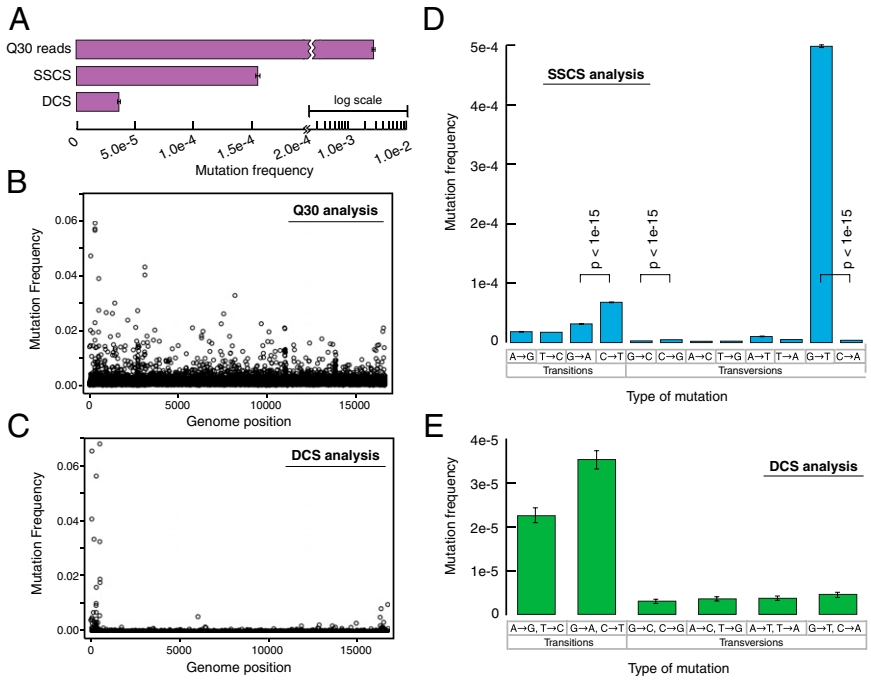

Fig. 4. Duplex Sequencing of human mitochondrial DNA. (*A*) Overall mutation frequency as measured by a standard sequencing approach, SSCS, and DCS. (*B*) Pattern of mutation in human mitochondrial DNA by a standard sequencing approach. The mutation frequency (vertical axis) is plotted for every position in the ∼16-kb mitochondrial genome. Due to the substantial background of technical error, no obvious mutational pattern is discernible by this method. (*C*) DCS analysis eliminates sequencing artifacts and reveals the true distribution of mitochondrial mutations to include a striking excess adjacent to the mtDNA origin of replication. (*D*) SSCS analysis yields a large excess of G→T mutations relative to complementary C→A mutations, consistent with artifacts from damaged-induced 8-oxo-G lesions during PCR. All significant ($P < 0.05$) differences between paired reciprocal mutation frequencies are noted. (*E*) DCS analysis removes the SSCS strand bias and reveals the true mtDNA mutational spectrum to be characterized by an excess of transitions.

sequences of tag-identified duplicates are compared such that mutations are scored only when present in multiple duplicates. This method, conceptually analogous to our approach of SSCS analysis, results in ∼20-fold improvement in accuracy relative to standard Illumina sequencing, but is presumably susceptible to the same sort of artifactual, largely damage-mediated, first-round PCR errors we observed in SSCS.

Notably, because SSCS is prone to damage-induced PCR errors, SSCS analysis can be used as a tool for detection of sites and patterns of DNA damage occurring in vivo. For example, the occurrence of G→T mutations in SSCS analysis in excess of reciprocal C→A mutations can be used as a marker for the extent of oxidative DNA damage in a sample. The ability to detect damage by SSCS could be further enhanced by using different DNA polymerases in the initial rounds of PCR, which have a proclivity to catalyze specific misinsertions opposite defined types of damage (51, 52).

In contrast to the damage sensitivity of single-strand consensus sequences, for DNA damage to result in an artifactual mutation in DCS, mutagenic lesions (or spontaneous, recurrent first-round PCR errors) would need to occur at the same nucleotide position on both strands of a molecule of duplex DNA *and* result in complementary errors. Thus, the background error frequency of our method may be calculated as (probability of error on one strand) × (probability of error on other strand) × (probability that both errors are complementary).

Based on the SSCS background error frequency of 3.4×10^{-5} from the M13mp2 DNA sequencing experiment, the error frequency of DCS can be approximated as: $(3.4 \times 10^{-5}) \times (3.4 \times 10^{-5}) \times {}^{1}/_{3} = 3.8 \times 10^{-10}$. This calculated error frequency represents a 10 million-fold improvement over the 3.8×10^{-3} value we obtained by standard methods. Of note, the calculation simplistically assumes that all mutational events are equally likely by multiplying by the factor one-third (because any given nucleotide can mutate to any one of three other nucleotides). In reality, the strong mutational bias observed in SSCSs indicates that reciprocal mispairs are not equally probable and, hence, the background of DCS is expected to be lower than this estimate.

In addition to their application for high sensitivity detection of rare DNA variants, the degenerate tags in our Duplex Sequencing adapters can also be used for single-molecule counting to precisely determine absolute DNA or RNA copy numbers (25, 29). Because tagging occurs before amplification, the relative abundance of variants in a population can be accurately assessed given that proportional representation is not subject to skewing by amplification biases. As with their use for error correction, because the degenerate tags are present in the adapters, there are no additional steps required during library preparation, which is in contrast to many existing methods of tag-based counting.

In principle, Duplex Sequencing could be performed on the Illumina or similar platforms without the use of Duplex Tags, but instead by using the randomly sheared ends of the DNA fragments as unique identifiers (20): specifically, for a given DNA sequence seen in sequencing read 1 with 5′ sheared end sequence α and 3′ sheared end sequence β used as a tag of form αβ, the partner strand will occur as a matching sequence in read 2 tagged with 5′ shear point β and 3′ shear point α. In practice, this approach will be limited by the finite number of possible shear points that overlap any given DNA position, and thus, will not be scalable to sequencing DNA at great depth at any given position. However, Duplex Sequencing analysis based on shear points alone may have a role for retrospective confirmation that specific mutations of interest are true mutations that were indeed present in the starting sample (i.e., present in both DNA strands), as opposed to technical artifacts. Overall, however, Duplex Sequencing is most generally applicable when randomized, complementary double-stranded tags are used. We used a 24-nucleotide tag in the current work, which yields up to $4^{24} = 2.8 \times 10^{14}$ distinct tag sequences. Combining information regarding the shear points of DNA with the tag sequence would allow a shorter tag to be used, thus minimizing loss of sequencing capacity owing to that used for sequencing of the tag sequence itself.

Once Duplex Tag-containing adapters are synthesized by a straightforward series of enzymatic steps, they can be substituted for standard sequencing adapters without any significant deviations from the normal workflow of sample preparation for Illumina sequencing. Moreover, Duplex Sequencing can be generalized to essentially any sequencing platform: a double-stranded tag can be incorporated into other existing adapters or for sequencing approaches that do not require adapters, a double-stranded tag can be ligated onto a duplex DNA sample before sequencing. In the setting of this simple compatibility with existing workflows, the ability of Duplex Sequencing to radically lower the error rate of sequencing, and the ability of the method to enable precise molecular counting and inference of sites of DNA damage, our

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1031.0005

Duplex Sequencing approach offers a powerful next-generation sequencing tool for diverse areas of medicine and biology.

Materials and Methods

Adapter Synthesis. Duplex Tag adapters were synthesized from two oligonucleotides with noncomplementary Y-shaped tails. A randomized single-stranded 12-nucleotide sequence present in one of the oligonucleotides was rendered double stranded by copying with DNA polymerase. After reaction cleanup, adapter A-tailing was performed by incubation with DNA polymerase and dATP. Further details are provided in *SI Materials and Methods*.

Sequencing Library Preparation. Double-stranded DNA was sheared and end repaired by standard protocols, followed by T-tailing with DNA polymerase and dTTP. T-tailed DNA was ligated to A-tailed Duplex Tag adapters, followed by PCR amplification for 18–20 cycles and sequencing on an Illumina HiSEq 2000. Further details are provided in *SI Materials and Methods*.

Data Processing. Reads were filtered for those containing a properly located tag sequence, and the 12-nucleotide tags present on each end of the paired reads were computationally combined to form a single 24-nucleotide tag for each read. Reads containing identical tag sequences were grouped together to form SSCS reads. Next, partner strands among SSCS reads were identified by virtue of the complementary tag sequences. The SSCS reads corresponding to both of the two strands of individual molecules of duplex DNA were then compared to form DCS reads. Resultant sequence positions were considered only when information from both DNA strands was in perfect agreement. Further details are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Elizabeth Swisher and Thomas Montine for providing human tissue samples and Tom Walsh, Anne Thornton, Ming Lee, and Bryan Paeper for assistance with DNA sequencing. This work was supported by National Institutes of Health Grants R01 CA115802, R01 CA102029, and P01 AG0751 (to L.A.L.); T32 AG000057 (to S.R.K.); F30 AG033485 (to J.J.S.); and F30 AG039173 (to J.B.H.).

1. Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135–1145.
2. Lecroq B, et al. (2011) Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc Natl Acad Sci USA* 108:13177–13182.
3. García-Garcerà M, et al. (2011) Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics. *PLoS ONE* 6:e24161.
4. Fordyce SL, et al. (2011) High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *Biotechniques* 51:127–133.
5. Druley TE, et al. (2009) Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 6:263–265.
6. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* 105:16266–16271.
7. Mitchell PS, et al. (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci USA* 105:10513–10518.
8. Boyd SD, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* 1(12):1–8.
9. Gundry M, Vijg J (2012) Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants. *Mutat Res* 729:1–15.
10. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* 96:317–323.
11. Goren A, et al. (2010) Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods* 7:47–49.
12. Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Res* 18:1687–1691.
13. Kozarewa I, et al. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 6:291–295.
14. Flaherty P, et al. (2012) Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res* 40:e2–e2.
15. Gerstung M, et al. (2012) Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* 3:811.
16. Campbell PJ, et al. (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci USA* 105:13081–13086.
17. Zagordi O, Klein R, Däumer M, Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 38:7400–7409.
18. Quail MA, et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.
19. Shen Y, et al. (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* 20:273–280.
20. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods* 7:119–122.
21. Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS (2004) Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res* 32:e135.
22. McCloskey ML, Stöger R, Hansen RS, Laird CD (2007) Encoding PCR products with batch-stamps and barcodes. *Biochem Genet* 45:761–767.
23. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
24. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci USA* 108:20166–20171.
25. Kivioja T, et al. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 9:72–74.
26. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res* 39:e81–e81.
27. Shiroguchi K, Jia TZ, Sims PA, Xie XS (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci USA* 109:1347–1352.
28. Fu GK, Hu J, Wang P-H, Fodor SPA (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci USA* 108:9026–9031.
29. Patwardhan RP, et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30:265–270.
30. Cervantes RB, Stringer JR, Shao C, Tischfield JA, Stambrook PJ (2002) Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proc Natl Acad Sci USA* 99:3586–3590.
31. Roach JC, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.
32. Lindahl T, Wood RD (1999) Quality control by DNA repair. *Science* 286:1897–1905.
33. Kunkel TA (1984) Mutational specificity of depurination. *Proc Natl Acad Sci USA* 81:1494–1498.
34. Thomas DC, et al. (1991) Fidelity of mammalian DNA replication and replicative DNA polymerases. *Biochemistry* 30:11751–11759.
35. Shibutani S, Takeshita M, Grollman AP (1991) Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* 349:431–434.
36. Kasai H, et al. (1993) Formation, inhibition of formation, and repair of oxidative 8-hydroxyguanine DNA damage. *Basic Life Sci* 61:257–262.
37. Stiller M, et al. (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proc Natl Acad Sci USA* 103:13578–13584.
38. McBride TJ, Preston BD, Loeb LA (1991) Mutagenic spectrum resulting from DNA damage by oxygen radicals. *Biochemistry* 30:207–213.
39. Greaves LC, et al. (2009) Quantification of mitochondrial DNA mutation load. *Aging Cell* 8:566–572.
40. Kraytsberg Y, Nicholas A, Caro P, Khrapko K (2008) Single molecule PCR in mtDNA mutational analysis: Genuine mutations vs. damage bypass-derived artifacts. *Methods* 46:269–273.
41. Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet* 59:501–509.
42. Parsons TJ, et al. (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet* 15:363–368.
43. Stoneking M (2000) Hypervariable sites in the mtDNA control region are mutational hotspots. *Am J Hum Genet* 67:1029–1032.
44. Kennedy SR, Loeb LA, Herr AJ (2011) Somatic mutations in aging, cancer and neurodegeneration. *Mech Ageing Dev*.
45. Vermulst M, et al. (2007) Mitochondrial point mutations do not limit the natural lifespan of mice. *Nat Genet* 39:540–543.
46. Song S, et al. (2005) DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proc Natl Acad Sci USA* 102:4990–4995.
47. Lin MT, Simon DK, Ahn CH, Kim LM, Beal MF (2002) High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain. *Hum Mol Genet* 11:133–145.
48. Jazin EE, Cavelier L, Eriksson I, Oreland L, Gyllensten U (1996) Human brain contains high levels of heteroplasmy in the noncoding regions of mitochondrial DNA. *Proc Natl Acad Sci USA* 93:12382–12387.
49. Khaidakov M, Heflich RH, Manjanatha MG, Myers MB, Aidoo A (2003) Accumulation of point mutations in mitochondrial DNA of aging mice. *Mutat Res* 526:1–7.
50. Haag-Liautard C, et al. (2008) Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. *PLoS Biol* 6:e204.
51. Schmitt MW, et al. (2010) Active site mutations in mammalian DNA polymerase delta alter accuracy and replication fork progression. *J Biol Chem* 285:32264–32272.
52. Niimi A, et al. (2004) Palm mutants in DNA polymerases alpha and eta alter DNA replication fidelity and translesion activity. *Mol Cell Biol* 24:2734–2746.

SEE COMMENTARY
GENETICS
Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1031.0006

Supporting Information

Supporting Information Corrected March 27, 2013

Schmitt et al. 10.1073/pnas.1208715109

SI Materials and Methods

Adapter Synthesis. Duplex Tag-labeled adapters were synthesized from two oligonucleotides (PAGE purified; Integrated DNA Technologies), designated as the primer strand: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT and the template strand: /5phos/ACTGNNNNNNNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. The two adapter strands were annealed by combining equimolar amounts of each oligo to a final concentration of 50 μM and heating to 95 °C for 5 min. The oligo mix was allowed to cool to room temperature over 1 h. The annealed primer-template complex was extended in a reaction consisting of 40 μM primer template, 25 units Klenow exo− DNA polymerase (New England Biolabs), 250 μM each dNTP, 50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, and 1 mM DTT for 1 h at 37 °C. The product was purified by ethanol precipitation. Due to the partial A-tailing property of Klenow exo-, this protocol results in a mixture of blunt-ended adapters and adapters with a single-nucleotide A overhang. A single-nucleotide A overhang was added to residual blunt fragments by incubating the adapters with 25 units Klenow exo-, 1 mM dATP, 50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, and 1 mM DTT for 1 h at 37 °C. The product was again ethanol precipitated and resuspended to a final concentration of 50 μM.

Construction of M13mp2 Variants. M13mp2 gapped DNA encoding the LacZ α fragment was extended by human DNA polymerase δ (1) and the resultant products were transformed into *Escherichia coli* and subjected to blue-white color screening as previously described (2). Mutant plaques were sequenced to determine the location of the mutation resulting in the color phenotype. A series of mutants, each differing from wild type by a single nucleotide change, were then mixed together with wild-type M13mp2 DNA to result in a single final mixture with distinct mutants represented at ratios of 1/10 (G6267A), 1/100 (T6299C), 1/1,000 (G6343A), and 1/10,000 (A6293T).

Oxidative Damage of M13mp2 DNA. Induction of DNA damage was performed by minor modifications to a published protocol (3): 300 ng of M13mp2 double-stranded DNA was incubated in 10 mM sodium phosphate buffer, pH 7.0, in the presence of 10 μM iron sulfate and 10 μM freshly diluted hydrogen peroxide. Incubation proceeded for 30 min at 37 °C in open 1.5-mL plastic microcentrifuge tubes.

DNA Isolation. M13mp2 DNA was isolated from *E. coli* strain MC1061 by Qiagen Miniprep. To allow for greater sequencing depth at a defined region of the M13mp2 genome, an 840-bp fragment was enriched by complete digestion with the restriction enzymes Bsu36I and NaeI (New England Biolabs), followed by isolation of the fragment on an agarose gel by the Recochip system (Takara Bio). Mitochondrial DNA was isolated as previously described (4).

Sequencing Library Preparation. A total of 3 μg of DNA was diluted into 130 μL of TE buffer (10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA) and was sheared on the Covaris AFA system with duty cycle 10%, intensity 5, cycles/burst 100, time 20 s × 6, temperature = 4 °C. DNA was purified with two volumes of Agencourt AMPure XP beads per manufacturer protocol. After end repair with the New England Biolab DNA End Repair kit per manufacturer 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 ∼200 bp or greater to bind to the beads). The beads were washed and DNA eluted. Standard Illumina library preparation protocols involve ligating A-tailed DNA to T-tailed adapters. However, as we used A-tailed adapters, the DNA was instead T-tailed. T-tailing was performed in a reaction containing 5 units Klenow exo-, 1 mM dTTP, 50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, and 1 mM DTT. The reaction proceeded for 1 h at 37 °C. DNA was purified with 1.2 volumes of AMPure XP beads. The custom Duplex Sequencing adapters were ligated by combining 750 ng of T-tailed DNA with 250 pmol adapters in a reaction containing 3,000 units T4 DNA ligase (Enzymatics), 50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT, and 1 mM ATP. The reaction was incubated at 25 °C for 15 min, and purified with 0.8 volumes of Ampure XP beads.

PCR Amplification and DNA Sequencing. Adapter-ligated DNA was amplified with the KAPA HiFi PCR kit (Kappa Biosciences) with PCR primers: AATGATACGGCGACCACCGAG and CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGC (where XXXXXX indicates the position of a fixed multiplexing barcode sequence). Following PCR amplification, the adapters contain all flow-cell and sequencing primer binding sites required for the Illumina TruSeq system. Duplex Sequencing is founded upon the concept of generating and sequencing multiple PCR duplicates of each strand of individual molecules of double-stranded DNA, thus the amount of input DNA and the number of PCR cycles need to be titrated to generate an average of at least three PCR duplicates per tag family. Excess PCR duplication, however, will result in unnecessary loss of sequencing capacity. We obtained adequate DNA duplication and reasonable sequencing capacity by amplifying 40 attomoles of adapter-ligated DNA for 18–20 cycles. DNA sequencing was then performed on the Illumina HiSEq 2000 system according to the manufacturer's recommendations.

Data Processing. Reads with intact Duplex Tags will consist of a 12-nucleotide random sequence, followed by a 5-nucleotide fixed sequence immediately upstream of captured DNA sequence. These reads were identified by filtering out reads that lack the expected fixed sequence at positions 13–17. The 12-nucleotide tag sequences from both the forward and reverse sequencing reads were computationally added to the read header to result in a combined 24-nt tag for each read, and the 5-nucleotide fixed sequence was removed. The first 4 nucleotides following the fixed adapter sequence were also removed to eliminate errors introduced during fragment end repair and ligation. Reads were then aligned to the reference genome with the Burrows-Wheeler aligner (BWA) and nonmapping reads were discarded. The entire human genome sequence (hg19) was used as reference for the mitochondrial DNA experiment, and reads that mapped to chromosomal DNA were removed. Reads sharing identical tag sequences were then grouped together and collapsed to consensus reads. Sequencing positions were discounted if the consensus group covering that position consisted of fewer than three members or if fewer than 90% of the sequences at that position in the consensus group had the identical sequence. A minimum group size of three was selected because next-generation se-

IDT Exhibit 1031.0007

quencing systems have an average base calling error rate of ~1/100. Requiring the same base to be identified in three distinct reads decreases the frequency of single-strand consensus sequence (SSCS) errors arising from base-call errors to $(1/100)^3 = 1 \times 10^{-6}$, which is below the frequency of spontaneous PCR errors that fundamentally limit the sensitivity of SSCSs. The requirement for 90% of sequences to agree to score a position is a highly conservative cutoff. For example, with a group size of eight, a single disagreeing read will lead to 87.5% agreement and the position will not be scored. If all groups in an experiment are of size nine or less, this cutoff will thus require perfect agreement at any given position to score the position. We anticipate that further development of our protocol may allow for less stringent parameters to be used to maximize the number of SSCS and duplex consensus sequence (DCS) reads that can be obtained from a given experiment.

Consensus reads were realigned with the BWA. The consensus sequences were then paired with their strand mate by grouping each 24-nucleotide tag of form αβ in read 1 with its corresponding tag of form βα 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 provided below.

Statistical Analysis. Ninety-five percent confidence intervals were determined with the Wilson score interval method. *P* values were calculated by the two-sample test for equality of proportions with continuity correction.

Overview of Duplex Sequencing Data Processing.

i) Discard reads that do not have the 5 nucleotide fixed sequence CAGTA present after exactly 12 random nucleotides, which comprise the Duplex Tag sequence.
ii) Combine the 12 nucleotide tags from read 1 and read 2 and transfer the combined 24-nucleotide tag sequence into the read header.
iii) Discard tags with inadequate complexity (i.e., those with >10 consecutive identical nucleotides).
iv) Remove the 5-nucleotide fixed sequence.
v) Trim an additional 4 nucleotides from the 5′ ends of each read pair (sites of error prone ligation and end repair).
vi) Align reads to the reference genome and discard nonmapping reads.
vii) Group together reads that have identical 24-nt tags, representing PCR duplicates of an individual single-stranded DNA fragment.
viii) Collapse tag families to SSCS reads, scoring only positions represented by three or more PCR duplicates and having >90% sequence identity among the duplicates.
ix) Realign reads to the reference genome.
x) For each read in read 1 file having tag sequence of format αβ, group with corresponding DCS partner in read 2 file with tag sequence of format βα.
xi) Only score positions with identical sequence among both DCS partners.

Example: Duplex Sequencing Tag Pairs. Consider the 4-nucleotide tags below, with flow cell sequences 1 and 2 in the locations marked and dashes representing a ligated DNA fragment. The Duplex Sequencing adapters actually contain 12-nucleotide Duplex Tags. Shorter tags are used here for clarity:

5′ 1-TAAC————TCCG-2 3′
3′ 2-ATTG————AGGC-1 5′.

The same molecules are shown again here, but with the lower strand now written in the 5′ → 3′ direction:

5′ 1-TAAC————TCCG-2 3′
5′ 1-CGGA————GTTA-2 3′.

These molecules are then PCR amplified and sequenced. They will yield the following reads:
the "top" strand:

5′ 1-TAAC————TCCG-2 3′

will give:

read 1 file: TAAC——
read 2 file: CGGA——.

Combining the read 1 and read 2 tags will produce the tag sequence:

TAACCGGA

the "bottom" strand:

5′ 1-CGGA————GTTA-2 3′

will give:

read 1 file: 1-CGGA——
read 2 file: 2-TAAC——.

Combining the read 1 and read 2 tags will produce the tag sequence:

CGGATAAC.

Note that the combined tags are of form αβ (read 1) and βα (read 2). The key concept is that read 2 is read by the sequencer as the complement of the strand containing read 1.

Example: Orientation of Paired Strand Mutations in Duplex Sequencing. In the initial DNA duplex shown above, now consider a mutation "x" paired to complementary nucleotide "y" that is on the "left" side of the DNA duplex:

1-TAAC—x——————TCCG-2
2-ATTG—y——————AGGC-1.

x will appear in read 1, and the complementary mutation on the opposite strand, y, will be seen in read 2. However, the mutation will appear specifically as x in both the read 1 and read 2 data, because y in read 2 is read out as x by the sequencer owing to the asymmetric nature of the sequencing primers, which generate the complementary sequence of the "lower" strand during read 2 as opposed to the direct sequence of the "top" strand during read 1.

If the identity of a base fails to match between the two reads, the position is considered undefined and is replaced by an "N" in the final sequence. For instance, with tag sequences denoted α and β, the sequence αβ-AACTGT in read 1 and βα-AAGTGT in read 2 would result in a final sequence of AANTGT.

1. Schmitt MW, Matsumoto Y, Loeb LA (2009) High fidelity and lesion bypass capability of human DNA polymerase delta. *Biochimie* 91:1163–1172.
2. Bebenek K, Kunkel TA (1995) Analyzing fidelity of DNA polymerases. *Methods Enzymol* 262:217–232.
3. McBride TJ, Preston BD, Loeb LA (1991) Mutagenic spectrum resulting from DNA damage by oxygen radicals. *Biochemistry* 30:207–213.
4. Vermulst M, et al. (2007) Mitochondrial point mutations do not limit the natural lifespan of mice. *Nat Genet* 39:540–543.

IDT Exhibit 1031.0008

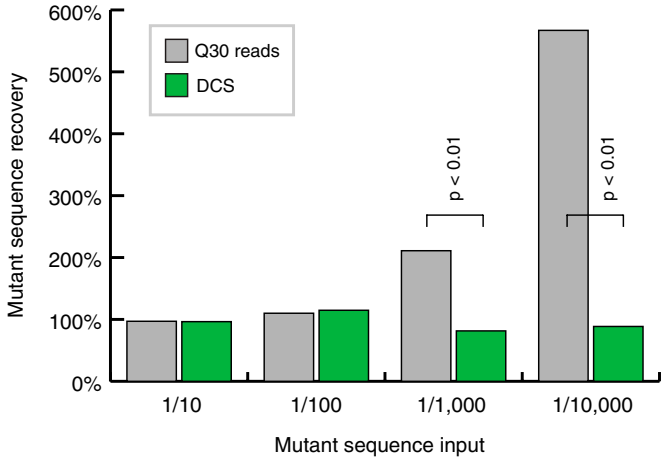

Fig. S1. Duplex Sequencing results in accurate recovery of spiked-in control mutations. A series of variants of M13mp2 DNA, each harboring a known single-nucleotide substitution, were mixed together at known ratios and the mixture was sequenced to ~20,000-fold final depth. Standard sequencing analysis cannot accurately distinguish mutants present at a ratio of less than 1/100, because artifactual mutations occurring at every position obscure the presence of less abundant true mutations, rendering apparent recovery greater than 100%. Duplex consensus sequences, in contrast, accurately identify spiked-in mutations down to the lowest tested ratio of 1/10,000.

Table S1. Data yield from Duplex Sequencing

	M13mp2 DNA	Mitochondrial DNA
Initial nucleotides	6.5×10^9	6.2×10^9
SSCS nucleotides	8.7×10^7	4.1×10^8
DCS nucleotides	$2.2E \times 10^7$	9.7×10^7
Initial reads per SSCS read	75	15
Initial reads per DCS read	295	64
SSCS reads per DCS read	4	4

Initial nucleotides represent raw reads that contain the expected fixed adapter sequence following 12 degenerate nucleotides and map to the reference genome. Apparent nucleotide loss in converting initial reads to SSCSs occurs because many of the initial reads intentionally represent identical PCR duplicates of single-stranded DNA molecules to allow for removal of sequencing and PCR errors by comparison of the sequence among the duplicates. A minimum of three initial reads are required to produce one SSCS; however, a greater average number is necessary to ensure that most DNA fragments have at least this number of duplicates. Under fully optimized conditions, each DCS read would arise from exactly two SSCS reads (one arising from each strand of the initial molecule of duplex DNA). An SSCS:DCS ratio greater than 2 indicates that the strand partner of some SSCSs was not recovered.

IDT Exhibit 1031.0009