# 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-

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.

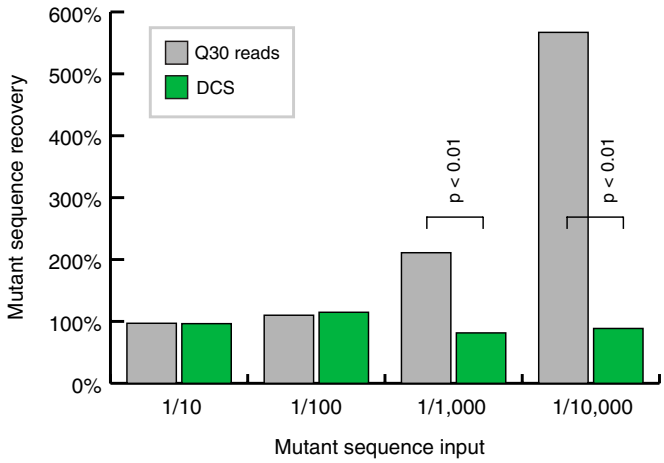


**Fig. S1.** Duplex Sequencing results in accurate recovery of spiked-in control mutations. A series of variants of M13mp2 DNA, each harboring a known single-nucleotide substitution, were mixed together at known ratios and the mixture was sequenced to ~20,000-fold final depth. Standard sequencing analysis 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 | $6.5 \times 10^9$ | $6.2 \times 10^9$ |
| SSCS nucleotides | $8.7 \times 10^7$ | $4.1 \times 10^8$ |
| DCS nucleotides | $2.2E \times 10^7$ | $9.7 \times 10^7$ |
| 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.