# Detection and quantification of rare mutations with massively parallel sequencing

**Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler[1], and Bert Vogelstein[1]**

The Ludwig Center for Cancer Genetics and Therapeutics and The Howard Hughes Medical Institute, Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231

Contributed by Bert Vogelstein, April 19, 2011 (sent for review March 21, 2011)

**The identification of mutations that are present in a small fraction of DNA templates is essential for progress in several areas of biomedical research. Although massively parallel sequencing instruments are in principle well suited to this task, the error rates in such instruments are generally too high to allow confident identification of rare variants. We here describe an approach that can substantially increase the sensitivity of massively parallel sequencing instruments for this purpose. The keys to this approach, called the Safe-Sequencing System ("Safe-SeqS"), are (*i*) assignment of a unique identifier (UID) to each template molecule, (*ii*) amplification of each uniquely tagged template molecule to create UID families, and (*iii*) redundant sequencing of the amplification products. PCR fragments with the same UID are considered mutant ("supermutants") only if ≥95% of them contain the identical mutation. We illustrate the utility of this approach for determining the fidelity of a polymerase, the accuracy of oligonucleotides synthesized in vitro, and the prevalence of mutations in the nuclear and mitochondrial genomes of normal cells.**

diagnostics | early diagnosis | biomarkers | genetics | cancer

Genetic mutations underlie many aspects of life and death—through evolution and disease, respectively. Accordingly, their measurement is critical to several fields of research. Luria and Delbrück's classic fluctuation analysis is a prototypic example of the insights into biological processes that can be gained simply by counting the number of mutations in carefully controlled experiments (1). Counting de novo mutations in humans, not present in their parents, has similarly led to new insights into the rate at which our species can evolve (2, 3). Similarly, counting genetic or epigenetic changes in tumors can inform fundamental issues in cancer biology (4). Mutations lie at the core of current problems in managing patients with viral diseases such as AIDS and hepatitis by virtue of the drug resistance they can cause (5, 6). Detection of such mutations, particularly at a stage before their becoming dominant in the population, will likely be essential to optimize therapy. Detection of donor DNA in the blood of organ transplant patients is an important indicator of graft rejection and detection of fetal DNA in maternal plasma can be used for prenatal diagnosis in a noninvasive fashion (7, 8). In neoplastic diseases, which are all driven by somatic mutations, the applications of rare mutant detection are manifold; they can be used to help identify residual disease at surgical margins or in lymph nodes, to follow the course of therapy when assessed in plasma, and to identify patients with early, surgically curable disease when evaluated in stool, sputum, plasma, and other bodily fluids (9–11).

These examples highlight the importance of identifying rare mutations for both basic and clinical research. Accordingly, innovative ways to assess them have been devised over the years. The first methods involved biologic assays based on prototrophy, resistance to viral infection or drugs, or biochemical assays (1, 12–18). Molecular cloning and sequencing provided a new dimension to the field, as they allowed the type of mutation, rather than simply its presence, to be identified (19–24). Some of the most powerful of these newer methods are based on digital PCR, in which individual molecules are assessed one by one (25). Digital PCR is conceptually identical to the analysis of individual clones of bacteria, cells, or virus, but is performed entirely in vitro with defined, inanimate reagents. Several implementations of digital PCR have been described, including the analysis of molecules arrayed in multiwell plates, in polonies, in microfluidic devices, and in water-in-oil emulsions (25–30). In each of these technologies, mutant templates are identified through their binding to oligonucleotides specific for the potentially mutant base.

Massively parallel sequencing represents a particularly powerful form of digital PCR in that hundreds of millions of template molecules can be analyzed one by one. It has the advantage over conventional digital PCR methods in that multiple bases can be queried sequentially and easily in an automated fashion. However, massively parallel sequencing cannot generally be used to detect rare variants because of the high error rate associated with the sequencing process. For example, with the commonly used Illumina sequencing instruments, this error rate varies from ~1% (31, 32) to ~0.05% (33, 34), depending on factors such as the read length (35), use of improved base-calling algorithms (36–38), and the type of variants detected (39). Some of these errors presumably result from mutations introduced during template preparation, during the preamplification steps required for library preparation, and during further solid-phase amplification on the instrument itself. Other errors are due to base misincorporation during sequencing and base-calling errors. Advances in base calling can enhance confidence (e.g., refs. 36–39), but instrument-based errors are still limiting, particularly in clinical samples wherein the mutation prevalence can be ≤0.01% (11). In the work described herein, we show how templates can be prepared and the sequencing data obtained from them more reliably interpreted, so that relatively rare mutations can be identified with commercially available instruments.

## Results

**Overview.** Our approach, called the Safe-Sequencing System ("Safe-SeqS"), involves two basic steps (Fig. 1). The first is the assignment of a unique identifier (UID) to each DNA template molecule to be analyzed. The second is the amplification of each uniquely tagged template, so that many daughter molecules with the identical sequence are generated (defined as a UID family). If a mutation preexisted in the template molecule used for amplification, that mutation should be present in every daughter molecule containing that UID (barring any subsequent replication or sequencing errors). A UID family in which at least 95% of family members have the identical mutation is called a "supermutant". Mutations not occurring in the original templates, such as those occurring during the amplification steps or through errors in base calling, should not give rise to supermutants. Conceptual and

Author contributions: I.K., N.P., K.W.K., and B.V. designed research; I.K., J.W., N.P., and B.V. performed research; I.K., J.W., N.P., K.W.K., and B.V. contributed new reagents/analytic tools; I.K., N.P., K.W.K., and B.V. analyzed data; and I.K. and B.V. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed: E-mail: kinzlke@jhmi.edu or bertvog@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1105422108/-/DCSupplemental.



EX1048

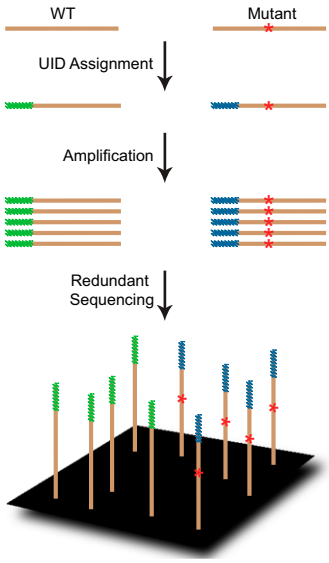


**Fig. 1.** Essential elements of Safe-SeqS. In the first step, each fragment to be analyzed is assigned a unique identification (UID) DNA sequence (green or blue bars). In the second step, the uniquely tagged fragments are amplified, producing UID families, each member of which has the same UID. A supermutant is defined as a UID family in which ≥95% of family members have the same mutation.

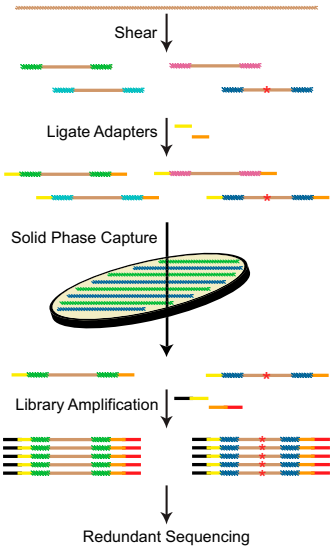


**Fig. 2.** Safe-SeqS with endogenous UIDs plus capture. The sequences of the ends of each fragment produced by random shearing (variously colored bars) serve as the unique identifiers (UIDs). These fragments are ligated to adapters (yellow and orange bars) so they can subsequently be amplified by PCR. One uniquely identifiable fragment is produced from each strand of the double-stranded template; only one strand is shown. Fragments of interest are captured on a solid phase containing oligonucleotides complementary to the sequences of interest. Following PCR amplification to produce UID families with primers containing 5′ "grafting" sequences (black and red bars), sequencing is performed and supermutants are defined as in Fig. 1.

practical issues related to UID assignment and supermutants are discussed in detail in *SI Materials and Methods*.

**Endogenous UIDs.** UIDs, sometimes called barcodes or indexes, can be assigned to nucleic acid fragments using a variety of methods. These methods include the introduction of exogenous sequences through PCR (40, 41) or ligation (42, 43). Even more simply, randomly sheared genomic DNA inherently contains UIDs consisting of the sequences of the two ends of each sheared fragment (Fig. 2 and Fig. S1). Paired-end sequencing of these fragments yields UID families that can be analyzed as described above. To use such endogenous UIDs in Safe-SeqS, we used two separate approaches: one designed to evaluate many genes simultaneously and the other designed to evaluate a single gene fragment in depth (Fig. 2 and Fig. S1, respectively).

For the evaluation of multiple genes, we ligated standard Illumina sequencing adapters to the ends of sheared DNA fragments to produce a standard sequencing library and then captured genes of interest on a solid phase (44). In this experiment, a library made from the DNA of ∼15,000 normal cells was used, and 2,594 bp from six genes were targeted for capture. After excluding known single-nucleotide polymorphisms, 25,563 apparent mutations, corresponding to $2.4 \times 10^{-4}$ mutations/bp, were also identified (Table 1). On the basis of previous analyses of mutation rates in human cells, at least 90% of these apparent mutations were likely to represent mutations introduced during template and library preparation or base-calling errors. Note that the error rate determined here ($2.4 \times 10^{-4}$ mutations/bp) is considerably lower than usually reported in experiments using the Illumina instrument because we used very stringent criteria for base calling (*SI Materials and Methods*).

With Safe-SeqS analysis of the same data, we determined that 69,505 original template molecules were assessed in this experiment (i.e., 69,505 UID families, with an average of 40 members per family, were identified) (Table 1). All of the polymorphic variants identified by conventional analysis were also identified by Safe-SeqS. However, only eight supermutants were observed among these families, corresponding to $3.5 \times 10^{-6}$ mutations/bp. Thus, Safe-SeqS decreased the presumptive sequencing errors by at least 70-fold.

A strategy using endogenous UIDs was also used to reduce false-positive mutations upon deep sequencing of a single region of interest. In this case, a library prepared as described above from ∼1,750 normal cells was used as template for inverse PCR using primers complementary to a gene of interest, so the PCR products could be directly used for sequencing (Fig. S1). With conventional analysis, an average of $2.3 \times 10^{-4}$ mutations/bp were observed, similar to that observed in the capture experiment (Table 1). Given that only 1,057 independent molecules from normal cells were assessed in this experiment, as determined through Safe-SeqS analysis, all mutations observed with conventional analysis likely represented false positives (Table 1). With Safe-SeqS analysis of the same data, no supermutants were identified at any position.

**Table 1. Safe-SeqS with endogenous UIDs**

| | Capture | Inverse PCR |
|---|---|---|
| Conventional analysis | | |
| High-quality base pairs | 106,958,863 | 1,041,346,645 |
| Mean high-quality base pairs read depth | 38,620× | 2,085,600× |
| Mutations identified | 25,563 | 234,352 |
| Mutations/bp | 2.4*E*-04 | 2.3*E*-04 |
| Safe-SeqS analysis | | |
| High-quality base pairs | 106,958,863 | 1,041,346,645 |
| Mean high-quality base pairs read depth | 38,620× | 2,085,600× |
| UID families | 69,505 | 1,057 |
| Average no. of members/UID family | 40 | 21,688 |
| Median no. of members/UID family | 19 | 4 |
| Supermutants identified | 8 | 0 |
| Supermutants/bp | 3.5*E*-06 | 0.0 |

GENETICS

**Exogenous UIDs.** Although the results described above show that Safe-SeqS can increase the reliability of massively parallel sequencing, the number of different molecules that can be examined using endogenous UIDs is limited. For fragments sheared to an average size of 150 bp (range 125–175), 36-base paired-end sequencing can evaluate a maximum of ~7,200 different molecules containing a specific mutation (2 reads × 2 orientations × 36 bases/read × 50-base variation on either end of the fragment). In practice, the actual number of UIDs is smaller because the shearing process is not entirely random.

To make more efficient use of the original templates, we developed a Safe-SeqS strategy that used a minimum number of enzymatic steps. This strategy also permitted the use of degraded or damaged DNA, such as found in clinical specimens or after bisulfite treatment for the examination of cytosine methylation (45). As depicted in Fig. 3, this strategy employs two sets of PCR primers. The first set is synthesized with standard phosphoramidite precursors and contained sequences complementary to the gene of interest on the 3′ end and different tails at the 5′ ends of both the forward and reverse primers. The different tails allowed universal amplification in the next step. Finally, there was a stretch of 12–14 random nucleotides between the tail and the sequence-specific nucleotides in the forward primer (40). The random nucleotides form the UIDs. An equivalent way to assign UIDs to fragments, not used in this study, would employ 10,000 forward primers and 10,000 reverse primers synthesized on a microarray. Each of these 20,000 primers would have gene-specific primers at their 3′ ends and one of 10,000 specific, predetermined, nonoverlapping UID sequences at their 5′ ends, allowing for $10^8$ [i.e., $(10^4)^2$] possible UID combinations. In either case, two cycles of PCR are performed with the primers and a high-fidelity polymerase, producing a uniquely tagged, double-stranded DNA fragment from each of the two strands of each original template molecule (Fig. 3). The residual, unused UID assignment primers are removed by digestion with a single strand-specific exonuclease, without further purification, and two new primers are added. The new primers, complementary to the tails introduced in the UID assignment cycles, contain grafting sequences at their 5′ ends, permitting solid-phase amplification on the Illumina instrument, and phosphorothioate residues at their 3′ ends to make them resistant to any remaining exonuclease. Following 25 additional cycles of PCR, the products are loaded on the Illumina instrument. As shown below, this strategy allowed us to evaluate the majority of input fragments and was used for several illustrative experiments.

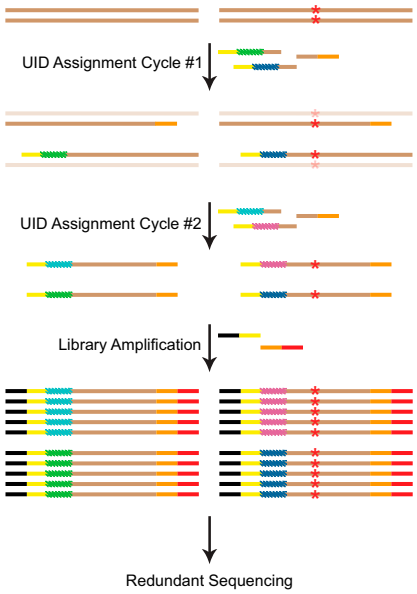


**Fig. 3.** Safe-SeqS with exogenous UIDs. DNA (sheared or unsheared) is amplified with a set of gene-specific primers. One of the primers has a random DNA sequence (e.g., a set of 14 Ns) that forms the unique identifier (UID) (variously colored bars), located 5′ to its gene-specific sequence, and both have sequences that permit universal amplification in the next step (yellow and orange bars). Two UID assignment cycles produce two fragments—each with a different UID—from each double-stranded template molecule, as shown. Subsequent PCR with universal primers, which also contain "grafting" sequences (black and red bars), produces UID families that are directly sequenced. Supermutants are defined as in the legend to Fig. 1.

**Analysis of DNA Polymerase Fidelity.** Measurement of the error rates of DNA polymerases is essential for their characterization and dictates the situations in which these enzymes can be used. We chose to measure the error rate of Phusion polymerase, as this polymerase has one of the lowest reported error frequencies of any commercially available enzyme and therefore poses a particular challenge for an in vitro-based approach. We first amplified a single human DNA template molecule, comprising a segment of an arbitrarily chosen human gene, through 19 rounds of PCR. The PCR products from these amplifications, in their entirety, were used as templates for Safe-SeqS as described in Fig. 3. In seven independent experiments of this type, the number of UID families identified by sequencing was 624,678 ± 421,274, which is consistent with an amplification efficiency of 92 ± 9.6% per round of PCR.

The error rate of Phusion polymerase, estimated through cloning of PCR products encoding β-galactosidase in plasmid vectors and transformation into bacteria, is reported by the manufacturer to be $4.4 \times 10^{-7}$errors/bp/PCR cycle. Even with very high-stringency base calling, conventional analysis of the Illumina sequencing data revealed an apparent error rate of $9.1 \times 10^{-6}$ errors/bp/PCR cycle, more than an order of magnitude higher than the reported Phusion polymerase error rate (Table 2, polymerase fidelity). In contrast, Safe-SeqS of the same data revealed an error rate of $4.5 \times 10^{-7}$errors/bp/PCR cycle, nearly identical to that measured for Phusion polymerase in biological assays (Table 2, polymerase fidelity). The vast majority (>99%) of these errors were single-base substitutions (Table S1, polymerase fidelity), consistent with previous data on the mutation spectra created by other prokaryotic DNA polymerases (15, 46, 47).

Safe-SeqS also allowed a determination of the total number of distinct mutational events and an estimation of PCR cycle in which the mutation occurred. There were 19 cycles of PCR performed in wells containing a single template molecule in these experiments. If a polymerase error occurred in cycle 19, there would be only one supermutant produced (from the strand containing the mutation). If the error occurred in cycle 18, there should be two supermutants (derived from the mutant strands produced in cycle 19), etc. Accordingly, the cycle in which the error occurred is related to the number of supermutants containing that error. The data from seven independent experiments demonstrate a relatively consistent number of observed total polymerase errors ($2.2 \pm 1.1 \times 10^{-6}$ distinct mutations/bp), in reasonable agreement with the number expected from simulations ($1.5 \pm 0.21 \times 10^{-6}$ distinct mutations/bp, detailed in *SI Materials and Methods*). The data also show a highly variable timing of occurrence of polymerase errors among experiments (Table S2), as predicted from classic fluctuation analysis (1). This kind of information is difficult to derive using conventional analysis of the same next-generation sequencing data, in part because of the prohibitively high apparent mutation rate noted above.

**Analysis of Oligonucleotide Composition.** A small number of mistakes during the synthesis of oligonucleotides from phoshoramidite precursors are tolerable for most applications, such as routine PCR or cloning. However, for synthetic biology, wherein many oligonucleotides must be joined together, such mistakes present a major

**Table 2. Safe-SeqS with exogenous UIDs**

| | Mean | SD |
|---|---|---|
| Polymerase fidelity | | |
| Conventional analysis of seven replicates | | |
| High-quality base pairs | 996,855,791 | 64,030,757 |
| Total mutations identified | 198,638 | 22,515 |
| Mutations/bp | 2.0*E*-04 | 1.7*E*-05 |
| Calculated Phusion error rate (errors/bp/cycle) | 9.1*E*-06 | 7.7*E*-07 |
| Safe-SeqS analysis of seven replicates | | |
| High-quality base pairs | 996,855,791 | 64,030,757 |
| UID families | 624,678 | 421,274 |
| Members/UID family | 107 | 122 |
| Total supermutants identified | 197 | 143 |
| Supermutants/bp | 9.9*E*-06 | 2.3*E*-06 |
| Calculated Phusion error rate (errors/bp/cycle) | 4.5*E*-07 | 1.0*E*-07 |
| *CTNNB1* mutations in DNA from normal human cells | | |
| Conventional analysis of three individuals | | |
| High-quality base pairs | 559,334,774 | 66,600,749 |
| Total mutations identified | 118,488 | 11,357 |
| Mutations/bp | 2.1*E*-04 | 1.6*E*-05 |
| Safe-SeqS analysis of three individuals | | |
| High-quality base pairs | 559,334,774 | 66,600,749 |
| UID families | 374,553 | 263,105 |
| Members/UID family | 68 | 38 |
| Total supermutants identified | 99 | 78 |
| Supermutants/bp | 9.0*E*-06 | 3.1*E*-06 |
| Mitochondrial mutations in DNA from normal human cells | | |
| Conventional analysis of seven individuals | | |
| High-quality base pairs | 147,673,456 | 54,308,546 |
| Total mutations identified | 30,599 | 12,970 |
| Mutations/bp | 2.1*E*-04 | 9.4*E*-05 |
| Safe-SeqS analysis of seven individuals | | |
| High-quality base pairs | 147,673,456 | 54,308,546 |
| UID families | 515,600 | 89,985 |
| Members/UID family | 15 | 6 |
| Total supermutants identified | 135 | 61 |
| Supermutants/bp | 1.4*E*-05 | 6.8*E*-06 |

obstacle to success. Clever strategies for making the gene construction process more efficient have been devised (48, 49), but all such strategies would benefit from more accurate synthesis of the oligonucleotides themselves. Determining the number of errors in synthesized oligonucleotides is difficult because the fraction of oligonucleotides containing errors can be lower than the sensitivity of conventional next-generation sequencing analyses.

To determine whether Safe-SeqS could be used for this determination, we used standard phosphoramidite chemistry to synthesize an oligonucleotide containing 31 bases that were designed to be identical to that analyzed in the polymerase fidelity experiment described above. In the synthetic oligonucleotide, the 31 bases were surrounded by sequences complementary to primers that could be used for the UID assignment steps of Safe-SeqS (Fig. 3). By performing Safe-SeqS on ~300,000 oligonucleotide templates, we found that there were $8.9 \pm 0.28 \times 10^{-4}$ supermutants/bp and that these errors occurred throughout the sequence of the oligonucleotides (Fig. S2*A*). The oligonucleotides contained a large number of insertion and deletion errors, representing $8.2 \pm 0.63\%$ and $25 \pm 1.5\%$ of the total supermutants, respectively. Importantly, both the position and the nature of the errors were highly reproducible among seven independent replicates of this experiment performed on the same batch of oligonucleotides (Fig. S2*A*). This nature and distribution of errors had little in common with that of the errors produced by Phusion polymerase (Fig. S2*B* and Table S3), which were distributed in the expected stochastic pattern among replicate experiments. The number of errors in the oligonucleotides synthesized with phosphoramidites was ~60 times higher than that in the equivalent products synthesized by Phusion polymerase. These data, in toto, indicate that the vast majority of errors in the former were generated during their synthesis rather than during the Safe-SeqS procedure.

Does Safe-SeqS preserve the ratio of mutant:normal sequences in the original templates? To address this question, we synthesized two 31-base oligonucleotides of identical sequence with the exception of nucleotide 15 (50:50 C/G instead of T) and mixed them at nominal mutant/normal fractions of 3.3% and 0.33%. Through Safe-SeqS analysis of the oligonucleotide mixtures, we found that the ratios were 2.8% and 0.27%, respectively. We conclude that the UID assignment and amplification procedures used in Safe-SeqS do not greatly alter the proportion of variant sequences and thereby provide a reliable estimate of that proportion when unknown. This conclusion is also supported by the reproducibility of variant fractions when analyzed in independent Safe-SeqS experiments (Fig. S2*A*).

**Analysis of DNA Sequences from Normal Human Cells.** The exogenous UID strategy (Fig. 3) was then used to determine the prevalence of rare mutations in a small region of the *CTNNB1* gene isolated from ~100,000 normal human cells from three unrelated individuals. Through comparison with the number of UID families obtained in the Safe-SeqS experiments (Table 2, *CTNNB1* mutations in DNA from normal human cells), we calculated that the majority ($78 \pm 9.8\%$) of the input fragments were converted into UID families. There was an average of 68 members/UID family, easily fulfilling the required redundancy for Safe-SeqS (Fig. S3). Conventional analysis of the Illumina sequencing data revealed an average of $118{,}488 \pm 11{,}357$ mutations among the ~560 Mb of sequence analyzed per sample, corresponding to an apparent mutation prevalence of $2.1 \pm 0.16 \times 10^{-4}$ mutations/bp (Table 2, *CTNNB1* mutations in DNA from normal human cells). Only an average of $99 \pm 78$ supermutants were observed in the Safe-SeqS analysis. The vast majority (>99%) of supermutants were single-base substitutions and the calculated mutation rate was $9.0 \pm 3.1 \times 10^{-6}$ mutations/bp (Table S1, *CTNNB1* mutations in DNA from normal human cells). Safe-SeqS thereby reduced the apparent frequency of mutations in genomic DNA by at least 24-fold (Fig. 4).

We applied the identical strategy to a short segment of mitochondrial DNA isolated from ~1,000 cells from each of seven unrelated individuals. Conventional analysis of the Illumina sequencing libraries produced with the Safe-SeqS procedure (Fig. 3) revealed an average of $30{,}599 \pm 12{,}970$ mutations among the ~150 Mb of sequence analyzed per sample, corresponding to an apparent mutation prevalence of $2.1 \pm 0.94 \times 10^{-4}$ mutations/bp (Table 2, mitochondrial mutations in DNA from normal human cells). Only $135 \pm 61$ supermutants were observed in the Safe-SeqS analysis. As with the *CTNNB1* gene, the vast majority of mutations were single-base substitutions, although occasional single-base deletions were also observed (Table S1, mitochondrial mutations in DNA from normal human cells). The calculated mutation rate in the analyzed segment of mtDNA was $1.4 \pm 0.68 \times 10^{-5}$ mutations/bp (Table 2, mitochondrial mutations in DNA from normal human cells). Thus, Safe-SeqS thereby reduced the apparent frequency of mutations in mitochondrial DNA by at least 15-fold.

## Discussion

The results described above demonstrate that the Safe-SeqS approach can substantially improve the accuracy of massively parallel sequencing (Tables 1 and 2). It can be implemented through either endogenous or exogenously introduced UIDs and can be applied to virtually any sample preparation workflow or sequencing platform. As demonstrated here, the approach can easily be used to identify rare mutants in a population of DNA templates, to measure poly-

GENETICS

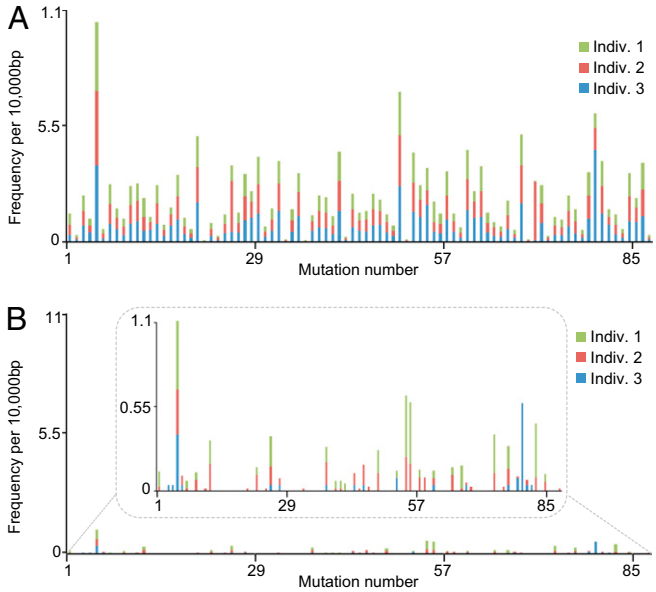


**Fig. 4.** Single-base substitutions identified by conventional and Safe-SeqS analysis. The exogenous UID strategy depicted in Fig. 3 was used to produce PCR fragments from the *CTNNB1* gene of three normal, unrelated individuals. Mutation numbers represent one of 87 possible single-base substitutions (3 possible substitutions/base × 29 bases analyzed). These fragments were sequenced on an Illumina GA IIx instrument and analyzed in the conventional manner (*A*) or with Safe-SeqS (*B*). Safe-SeqS results are displayed on the same scale as conventional analysis for direct comparison; the *Inset* is a magnified view. Note that most of the variants identified by conventional analysis are likely to represent sequencing errors, as indicated by their high frequency relative to Safe-SeqS and their consistency among unrelated samples.

merase error rates, and to judge the reliability of oligonucleotide syntheses. One of the advantages of the strategy is that it yields the number of templates analyzed as well as the fraction of templates containing variant bases. Previously described in vitro methods for the detection of small numbers of template molecules (e.g., refs. 29 and 50) allow the fraction of mutant templates to be determined but cannot determine the number of mutant and normal templates in the original sample.

It is of interest to compare Safe-SeqS to other approaches for reducing errors in next-generation sequencing. As mentioned in the Introduction, sophisticated algorithms to increase the accuracy of base calling have been developed (e.g., refs. 36–39). These improved base calling algorithms can certainly reduce false-positive calls, but their effectiveness is still limited by artifactual mutations occurring during the PCR steps required for library preparation as well as by any residual base-calling errors. For example, the algorithm used in the current study used very stringent criteria for base calling and was applied to short read lengths, but was still unable to reduce the error rate to less than an average of $2.0 \times 10^{-4}$ errors/bp. This error frequency is at least as low as those reported with other algorithms. To improve sensitivity further, these base-calling improvements can be used together with Safe-SeqS. Travers et al. describe another powerful strategy for reducing errors (51). With this technology, both strands of each template molecule are sequenced redundantly after a number of preparative enzymatic steps. However, this approach can be performed only on a specific instrument. Moreover, for many clinical applications, there are relatively few template molecules in the initial sample and evaluation of nearly all of them is required to obtain the requisite sensitivity. The approach described here with exogenously introduced UIDs (Fig. 3) fulfills this requirement by coupling the UID assignment step with a subsequent amplification in which few molecules are lost. Our endogenous UID approaches (Fig. 2 and Fig. S1) and the one described by Travers et al. are not ideally suited for this purpose because of the inevitable losses of template molecules during the ligation and other preparative steps.

How do we know that the mutations identified by conventional analyses in the current study represent artifacts rather than true mutations in the original templates? Strong evidence supporting this is provided by the observation that the mutation prevalence in all but one experiment was similar: $2.0 \times 10^{-4}$–$2.4 \times 10^{-4}$ mutations/bp (Tables 1 and 2). The exception was the experiment with oligonucleotides synthesized from phosphoramidites, in which the error of the synthetic process was apparently higher than the error rate of conventional Illumina analysis when used with stringent base-calling criteria. In contrast, the mutation prevalence of Safe-SeqS varied much more, from 0.0 to $1.4 \times 10^{-5}$ mutations/bp, depending on the template and experiment. Moreover, the mutation prevalence measured by Safe-SeqS in the most controlled experiment, in which polymerase fidelity was measured (Table 2, polymerase fidelity), was almost identical to that predicted from previous experiments in which polymerase fidelity was measured by biological assays. Our measurements of mutation prevalence in the DNA from normal cells are consistent with some previous experimental data. However, estimates of these prevalences vary widely and may depend on cell type and sequence analyzed (*SI Materials and Methods*). We therefore cannot be certain that the relatively low number of mutations revealed by Safe-SeqS represented errors occurring during the sequencing process rather than true mutations present in the original DNA templates. Potential sources of error in the Safe-SeqS process are described in *SI Materials and Methods*.

Like all techniques, Safe-SeqS has limitations. For example, we have demonstrated that the exogenous UIDs strategy can be used to analyze a single amplicon in depth. This technology may not be applicable to situations wherein multiple amplicons must be analyzed from a sample containing a limited number of templates. Multiplexing in the UID assignment cycles (Fig. 3) may provide a solution to this challenge. A second limitation is that the efficiency of amplification in the UID assignment cycles is critical for the success of the method. Clinical samples can contain inhibitors that reduce the efficiency of this step. This problem can presumably be overcome by performing more than two cycles in the UID assignment PCR step (Fig. 3), although this would complicate the determination of the number of templates analyzed. The specificity of Safe-SeqS is currently limited by the fidelity of the polymerase used in the UID assignment PCR step, i.e., $8.8 \times 10^{-7}$ mutations/bp in its current implementation with two cycles. Increasing the number of cycles in the UID assignment PCR step to five would decrease the overall specificity to $\sim 2 \times 10^{-6}$ mutations/bp. However, this specificity can be increased by requiring more than one supermutant for mutation identification—the probability of introducing the same artifactual mutation twice or three times would be exceedingly low [$(2 \times 10^{-6})^2$ or $(2 \times 10^{-6})^3$, respectively]. In sum, there are several simple ways to vary the Safe-SeqS procedure and analysis to realize the needs of specific experiments.

Luria and Delbrück, in their classic paper in 1943, wrote that their "prediction cannot be verified directly, because what we observe, when we count the number of resistant bacteria in a culture, is not the number of mutations which have occurred but the number of resistant bacteria which have arisen by multiplication of those which mutated, the amount of multiplication depending on how far back the mutation occurred" (ref. 1, p. 495). The Safe-SeqS procedure described here can verify such predictions because the number as well as the time of occurrence of each mutation can be estimated from the data, as noted in the experiments on polymerase fidelity. In addition to templates generated by polymerases in vitro, the same approach can be applied to DNA from bacteria, viruses, and mammalian cells. We therefore expect that this

strategy will provide definitive answers to a variety of important biomedical questions.

## Materials and Methods

**Endogenous UIDs.** To expose endogenous UIDs, DNA was fragmented to an average size of ∼200 bp by acoustic shearing (Covaris) and then end-repaired, A-tailed, and ligated to Y-shaped adapters according to standard Illumina protocols. DNA was captured (44) with a filter containing 2,594 nt corresponding to six cancer genes. For the inverse PCR experiments (Fig. S1), we ligated custom adapters (IDT) (Table S4) instead of standard Y-shaped Illumina adapters to sheared cellular DNA. Inverse PCR was performed using *KRAS* forward and reverse primers (Table S4) that both contained grafting sequences for hybridization to the Illumina GA IIx flow cell (Table S4). Further details are provided in *SI Materials and Methods*.

**Exogenous UIDs.** Each strand of each template molecule was encoded with a 12- or 14-base UID using two cycles of amplicon-specific PCR, as described in the text and Fig. 3. The amplicon-specific primers both contained universal tag sequences at their 5′ ends for a later amplification step. The UIDs constituted 12 or 14 random nucleotide sequences appended to the 5′ end of the forward amplicon-specific primers (Table S4). Following two cycles of PCR for UID assignment, the products were digested with a single-strand DNA-specific nuclease. Primers complementary to the introduced universal tags and containing 3′-terminal phosphorothioates (Table S4) were added and 25 additional cycles of PCR were performed. Further details are provided in *SI Materials and Methods*.

**Sequencing.** Sequencing of all of the libraries described above was performed using an Illumina GA IIx instrument as specified by the manufacturer. High-quality reads were grouped into UID families on the basis of their endogenous or exogenous UIDs. Only UID families with two or more members were considered, as described in detail in *SI Materials and Methods*.

**ACKNOWLEDGMENTS.** This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research; The Virginia and D. K. Ludwig Fund for Cancer Research; The Sol Goldman Center for Pancreatic Cancer Research; The Joseph L. Rabinowitz Fund for Pancreatic Cancer Research; National Cancer Institute Division of Cancer Prevention Contract N01-CN-43302; and National Institutes of Health Grants CA62924, CA43460, and CA57345. I.K. is also supported by a United Negro College Fund–Merck Graduate Fellowship.

1. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.
2. Roach JC, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.
3. Durbin RM, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
4. Shibata D (2011) Mutation and epigenetic molecular clocks in cancer. *Carcinogenesis* 32:123–128.
5. McMahon MA, et al. (2007) The HBV drug entecavir - effects on HIV-1 replication and resistance. *N Engl J Med* 356:2614–2621.
6. Eastman PS, et al. (1998) Maternal viral genotypic zidovudine resistance and infrequent failure of zidovudine therapy to prevent perinatal transmission of human immunodeficiency virus type 1 in pediatric AIDS Clinical Trials Group Protocol 076. *J Infect Dis* 177:557–564.
7. Chiu RW, et al. (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci USA* 105:20458–20463.
8. 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.
9. Hoque MO, et al. (2003) High-throughput molecular analysis of urine sediment for the detection of bladder cancer by high-density single-nucleotide polymorphism array. *Cancer Res* 63:5723–5726.
10. Thunnissen FB (2003) Sputum examination for early detection of lung cancer. *J Clin Pathol* 56:805–810.
11. Diehl F, et al. (2008) Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. *Gastroenterology* 135:489–498.
12. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* 112:29–35.
13. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.
14. Campbell F, Appleton MA, Shields CJ, Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right- and left-sided sporadic colorectal carcinomas. *J Pathol* 186:31–35.
15. Tindall KR, Kunkel TA (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry* 27:6008–6013.
16. Kunkel TA (1985) The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol Chem* 260:5787–5796.
17. van Dongen JJ, Wolvers-Tettero IL (1991) Analysis of immunoglobulin and T cell receptor genes. Part II: Possibilities and limitations in the diagnosis and management of lymphoproliferative diseases and related disorders. *Clin Chim Acta* 198:93–174.
18. Grist SA, McCarron M, Kutlaca A, Turner DR, Morley AA (1992) In vivo human somatic mutation: Frequency and spectrum with age. *Mutat Res* 266:189–196.
19. Liu Q, Sommer SS (2004) Detection of extremely rare alleles by bidirectional pyrophosphorolysis-activated polymerization allele-specific amplification (Bi-PAP-A): Measurement of mutation load in mammalian tissues. *Biotechniques* 36:156–166.
20. Monnat RJ, Jr., Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. *Proc Natl Acad Sci USA* 82:2895–2899.
21. Shi C, et al. (2004) LigAmp for sensitive detection of single-nucleotide differences. *Nat Methods* 1:141–147.
22. Keohavong P, Thilly WG (1989) Fidelity of DNA polymerases in DNA amplification. *Proc Natl Acad Sci USA* 86:9253–9257.
23. Sidransky D, et al. (1991) Identification of p53 gene mutations in bladder cancers and urine samples. *Science* 252:706–709.
24. Bielas JH, Loeb LA (2005) Quantification of random genomic mutations. *Nat Methods* 2:285–290.
25. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
26. Mitra RD, et al. (2003) Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci USA* 100:5926–5931.
27. Chetverina HV, Samatov TR, Ugarov VI, Chetverin AB (2002) Molecular colony diagnostics: Detection and quantitation of viral nucleic acids by in-gel PCR. *Biotechniques* 33:150–152, 154, 156.
28. Zimmermann BG, et al. (2008) Digital PCR: A powerful new tool for noninvasive prenatal diagnosis? *Prenat Diagn* 28:1087–1093.
29. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* 100:8817–8822.
30. Ottesen EA, Hong JW, Quake SR, Leadbetter JR (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* 314:1464–1467.
31. Quail MA, et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.
32. Nazarian R, et al. (2010) Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* 468:973–977.
33. He Y, et al. (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464:610–614.
34. Gore A, et al. (2011) Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63–67.
35. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.
36. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-Cyclic: A self-optimizing base caller for next-generation sequencing. *Nat Methods* 5:679–682.
37. Rougemont J, et al. (2008) Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 9:431.
38. Druley TE, et al. (2009) Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 6:263–265.
39. Vallania FL, et al. (2010) High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res* 20:1711–1718.
40. McCloskey ML, Stöger R, Hansen RS, Laird CD (2007) Encoding PCR products with batch-stamps and barcodes. *Biochem Genet* 45:761–767.
41. Parameswaran P, et al. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 35:e130.
42. Craig DW, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* 5:887–893.
43. 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.
44. Herman DS, et al. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* 6:507–510.
45. Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128:683–692.
46. de Boer JG, Ripley LS (1988) An in vitro assay for frameshift mutations: Hotspots for deletions of 1 bp by Klenow-fragment polymerase share a consensus DNA sequence. *Genetics* 118:181–191.
47. Eckert KA, Kunkel TA (1990) High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Res* 18:3739–3744.
48. Kosuri S, et al. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* 28:1295–1299.
49. Matzas M, et al. (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat Biotechnol* 28:1291–1294.
50. Li J, et al. (2008) Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nat Med* 14:579–584.
51. Eid J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

# Supporting Information

## Kinde et al. 10.1073/pnas.1105422108

### SI Materials and Methods

**Endogenous UIDs.** Genomic DNA from human pancreas or cultured lymphoblastoid cells was prepared using Qiagen kits. The pancreas DNA was used for the capture experiment and the lymphoblastoid cells were used for the inverse PCR experiment. DNA was quantified by optical absorbance and with qPCR. DNA was fragmented to an average size of ∼200 bp by acoustic shearing (Covaris) and then end-repaired, A-tailed, and ligated to Y-shaped adapters according to standard Illumina protocols. The ends of each template molecule provide endogenous UIDs corresponding to their chromosomal positions. After PCR-mediated amplification of the libraries with primer sequences within the adapters, DNA was captured (1) with a filter containing 2,594 nt corresponding to six cancer genes. After capture, 18 cycles of PCR were performed to ensure sufficient amounts of template for sequencing on an Illumina GA IIx instrument.

For the inverse PCR experiments (Fig. S1), we ligated custom adapters (IDT) (Table S4) instead of standard Y-shaped Illumina adapters to sheared cellular DNA. These adapters retained the region complementary to the universal sequencing primer but lacked the grafting sequences required for hybridization to the Illumina GA IIx flow cell. The ligated DNA was diluted into 96 wells and the DNA in each column of 8 wells was amplified with a unique forward primer containing one of 12 index sequences at its 5′ end plus a standard reverse primer (Table S4). Amplifications were performed using Phusion HotStart I (NEB) in 50-μL reactions containing 1× Phusion HF buffer, 0.5 mM dNTPs, 0.5 μM each forward and reverse primer (both 5′-phosphorylated), and 1 unit of Phusion polymerase. The following cycling conditions were used: 1 cycle of 98 °C for 30 s and 16 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s. All 96 reactions were pooled and then purified using a Qiagen MinElute PCR Purification kit (no. 28004) and a QIAquick Gel Extraction kit (no. 28704). To prepare the circular templates necessary for inverse PCR, DNA was diluted to ∼1 ng/uL and ligated with T4 DNA Ligase (Enzymatics) for 30 min at room temperature in a 600-uL reaction containing 1× T4 DNA ligation buffer and 18,000 units of T4 DNA Ligase. The ligation reaction was purified using a Qiagen MinElute kit. Inverse PCR was performed using Phusion Hot Start I on 90 ng of circular template distributed in twelve 50-μL reactions, each containing 1× Phusion HF buffer, 0.25 mM dNTPs, 0.5 uM each of *KRAS* forward and reverse primers (Table S4), and 1 unit of Phusion polymerase. The *KRAS*-specific primers both contained grafting sequences for hybridization to the Illumina GA IIx flow cell (Table S4). The following cycling conditions were used: 1 cycle of 98 °C for 2 min and then 37 cycles of 98 °C for 10 s, 61 °C for 15 s, and 72 °C for 10 s. The final purification was performed with a NucleoSpin Extract II kit (Macherey-Nagel) and eluted in 20 μL NE buffer. The resulting DNA fragments contained UIDs composed of three sequences: 2 endogenous ones, represented by the two ends of the original sheared fragments, plus the exogenous sequence introduced during the indexing amplification. As 12 exogenous sequences were used, this increased the number of distinct UIDs by 12-fold over that obtained without exogenous UIDs. This number could easily be increased by using a greater number of indexing primers.

**Exogenous UIDs.** Genomic DNA from normal human colonic mucosae or blood lymphocytes was prepared using Qiagen kits. The DNA from colonic mucosae was used for the experiments on *CTNNB1* and mitochondrial DNA, whereas the lymphocyte DNA was used for the experiments on *CTNNB1* and on polymerase fidelity. DNA was quantified with digital PCR (2) using primers that amplified single-copy genes from human cells (analysis of polymerase fidelity in *Analysis of DNA Polymerase Fidelity* and *CTNNB1* in *Analysis of DNA Sequences from Normal Human Cells*), with qPCR (mitochondrial DNA), or by optical absorbance (oligonucleotides in *Analysis of Oligonucleotide Composition*). Each strand of each template molecule was encoded with a 12- or 14-base UID using two cycles of amplicon-specific PCR, as described in the text and Fig. 3. The amplicon-specific primers both contained universal tag sequences at their 5′ ends for a later amplification step. The UIDs constituted 12 or 14 random nucleotide sequences appended to the 5′ end of the forward amplicon-specific primers (Table S4). These primers can generate 16.8 and 268 million distinct UIDs, respectively. It is important that the number of distinct UIDs greatly exceeds the number of original template molecules to minimize the probability that two different original templates acquire the same UID. The UID assignment PCR cycles included Phusion Hot Start II (NEB) in a 45-μL reaction containing 1× Phusion HF buffer, 0.25 mM dNTPs, 0.5 μM each forward (containing 12–14 Ns) and reverse primers, and 2 units of Phusion polymerase. To keep the final template concentrations <1.5 ng/uL, multiple wells were used to create some libraries. The following cycling conditions were used: one cycle of 98 °C for 30 s and then two cycles of 98 °C for 10 s, 61 °C for 120 s, and 72 °C for 10 s. To ensure complete removal of the first-round primers, each well was digested with 60 units of a single-strand DNA-specific nuclease (Exonuclease-I; Enzymatics) at 37 °C for 1 h. After a 5-min heat inactivation at 98 °C, primers complementary to the introduced universal tags (Table S4) were added to a final concentration of 0.5 uM each. These primers contained two terminal phosphorothioates to make them resistant to any residual Exonuclease-I activity. They also contained 5′-grafting sequences necessary for hybridization to the Illumina GA IIx flow cell. Finally, they contained an index sequence between the grafting sequence and the universal tag sequence. This index sequence enables the PCR products from multiple different individuals to be simultaneously analyzed in the same flow cell compartment of the sequencer. The following cycling conditions were used for the subsequent 25 cycles of PCR: 98 °C for 10 s and 72 °C for 15 s. No intermediate purification steps were performed in an effort to reduce the losses of template molecules.

After the second round of amplification, wells were consolidated and purified using a Qiagen QIAquick PCR Purification kit (no. 28104) and eluted in 50 μL EB buffer (Qiagen). Fragments of the expected size were purified after agarose (mtDNA libraries) or polyacrylamide (all other libraries) gel electrophoresis. For agarose gel purification, the eight 6-μL aliquots were loaded into wells of a 2% Size Select gel (Invitrogen) and bands of the expected size were collected in EB buffer as specified by the manufacturer. For polyacrylamide gel purification, ten 5-μL aliquots were loaded into wells of a 10% TBE Polyacrylamide gel (Invitrogen). Gel slices containing the fragments of interest were excised, crushed, and eluted essentially as described (3).

**Analysis of Phusion Polymerase Fidelity.** Amplification of a fragment of human genomic DNA within the BMX (RefSeq accession NM_203281.2) gene was first performed using the PCR conditions described above. The template was diluted so that an average of one template molecule was present in every three wells of a 96-well PCR plate. Fifty-microliter PCR reactions were then performed in 1× Phusion HF buffer, 0.25 mM dNTPs, 0.5 μM each forward and reverse primers (Table S4), and 2 units of

Phusion polymerase. The cycling conditions were 1 cycle of 98 °C for 30 s and 19 cycles of 98 °C for 10 s, 61 °C for 120 s, and 72 °C for 10 s. The primers were removed by digestion with 60 units of Exonuclease-I at 37 °C for 1 h followed by a 5-min heat inactivation at 98 °C. No purification of the PCR product was performed, either before or after Exonuclease-I digestion. The entire contents of each well were then used as templates for the exogenous UIDs strategy described above.

**Sequencing.** Sequencing of all of the libraries described above was performed using an Illumina GA IIx instrument as specified by the manufacturer. The total length of the reads used for each experiment varied from 36 to 73 bases. Base calling and sequence alignment were performed with the Eland pipeline (Illumina). Only high-quality reads meeting the following criteria were used for subsequent analysis: (*i*) the first 25 bases passed the standard Illumina chastity filter; (*ii*) every base in the read had a quality score ≥20; and (*iii*) ≤3 mismatches to expected sequences. For the exogenous UID libraries, we additionally required the UIDs to have a quality score ≥30. We noticed a relatively high frequency of errors at the ends of the reads in the endogenous UID libraries prepared with the standard Illumina protocol, presumably introduced during shearing or end repair, so the first and last 3 bases of these tags were excluded from analysis.

**Safe-SeqS Analysis.** High-quality reads were grouped into UID families on the basis of their endogenous or exogenous UIDs. Only UID families with two or more members were considered. Such UID families included the vast majority (≥99%) of the sequencing reads. To ensure that the same data were used for both conventional and Safe-SeqS analysis, we also excluded UID families containing only one member from conventional analysis. Furthermore, we identified a base as mutant in conventional sequencing analysis only if the same variant was identified in at least two members of at least one UID family (i.e., two mutations) when comparing conventional analysis to that of Safe-SeqS with exogenous UIDs. For comparison with Safe-SeqS with endogenous UIDs, we required at least two members of each of two UID families (i.e., four mutations) to identify a position as mutant in conventional analysis. With either endogenous or exogenous UIDs, a supermutant was defined as a UID family in which ≥95% of members shared the identical mutation. Thus, UID families with <20 members had to be 100% identical at the mutant position, whereas a 5% combined replication and sequencing error rate was permitted in UID families with more members. To determine polymerase fidelity using Safe-SeqS, and to compare the results with previous analyses of Phusion polymerase fidelity, it is important to realize that the previous analyses would detect only mutations present in both strands of the PCR products (4). This would be equivalent to analyzing PCR products generated with one less cycle with Safe-SeqS, and the appropriate correction was made in Table 2 (polymerase fidelity). Unless otherwise specified, all values listed in the text and tables represent means and SDs.

***Error-generating processes.*** Apparent mutations, defined as any base call that varies from the expected base at a defined position, can result from a variety of processes:

*i*) Mutations present in the template DNA. For templates derived from normal human cells, these include mutations that were present in the zygote, occurred later during embryonic and adult development, or were present in a contaminant inadvertently introduced into the sample. These mutations are expected to be present in both strands of the relevant templates. If the mutation occurred only in the last cell cycle of a cell whose DNA was used as template, the mutation would be present in only one strand of the template.

*ii*) Chemically modified bases present in the templates. It has been estimated that there are many thousands of oxidized bases present in every human cell (5). When such DNA is amplified by Phusion polymerase, an apparent mutation in one strand may result.

*iii*) Errors introduced during the shearing process required to generate small fragments for sequencing. Acoustic shearing generates short-lived, high temperatures that can damage DNA.

*iv*) Errors introduced during end repair of the sheared fragments. The source of these errors can be polymerase infidelity or through incorporation of chemically modified bases in the dNTPs used for polymerization.

*v*) Errors introduced by other enzymatic steps, particularly if the enzymes are impure and contaminated with nucleases, polymerases, or ligases.

*vi*) Errors introduced during PCR amplification to prepare the libraries for capturing or for inverse PCR.

*vii*) Errors during PCR after capturing or during inverse PCR amplification.

*viii*) Errors introduced into the UID assignment cycles of Safe-SeqS (Fig. 3).

*ix*) Errors introduced into the library amplification cycles of Safe-SeqS performed with exogenous UIDs. Note that if UID assignment primers from process *viii* are not completely removed, they could potentially amplify DNA fragments containing errors introduced during these cycles, creating a new supermutant.

*x*) Errors introduced into the first bridge-PCR cycle on the Illumina flow cell. If amplification is inefficient, an error introduced into the second bridge-PCR cycle could also result in a cluster containing a mutation in most of its component molecules.

*xi*) Errors in base calling.

***Achieving accuracy with Safe-SeqS.*** With conventional sequencing-by-synthesis approaches, all of the error-producing processes described above are relevant, resulting in a relatively high number of false-positive mutation calls (Tables 1 and 2). Safe-SeqS minimizes the number of false-positive mutation calls in several ways. Safe-SeqS with exogenous UIDs results in the fewest false-positive mutation calls because it requires the fewest enzymatic steps. With exogenous UIDs, error-generating processes *iii–vii* are completely eliminated because these steps are not performed. Safe-SeqS with exogenous UIDs also drastically reduces errors resulting from error-generating processes *x* and *xi* because of the way the data are analyzed.

After Safe-SeqS with exogenous UIDs, the only false-positive errors remaining should be those introduced during the UID assignment PCR cycles (error-generating process *viii*) or residual UID-containing primers during the library amplification cycles (error-generating process *ix*). The errors from error-generating process *viii* can be reduced by requiring at least two supermutants to identify a position as mutant. This requirement is reasonable because every preexisting mutation in a double-stranded DNA template should give rise to two supermutants, one from each strand. Furthermore, this requirement would reduce error-generating process *ii* (damaged bases in the original templates) because such bases, when copied, should give rise to only one supermutant. Finally, errors generated during the library amplification cycles (process *ix*) will not be amplified by residual UID-containing primers if those primers are completely removed, such as attempted in our study with excess Exonuclease-I.

With endogenous UIDs, the mistakes introduced by processes *x* and *xi* are drastically reduced because of the way in which the data are analyzed (as with exogenous UIDs). Errors introduced in processes *ii–vii* can be minimized by requiring that a mutation be observed in at least two UID families, for the reasons stated in the paragraph above. With this requirement, few false-positive mutations, in theory, should be identified.

In practice, the situation is complicated by the fact that the various amplifications are not perfect, so every strand of every original template molecule is not recovered as a UID family. This efficiency can vary from sample to sample, depending in part on the concentration of inhibitors present in clinical samples. Moreover, with exogenous UIDs, a polymerase error during the library amplification step can create a new UID family that was not represented in the UID assignment step. If this error occurred in a mutant template, an additional, artificial supermutant would be created.

These factors can be managed by incorporating various additional criteria into the analyses. For example, one might require UID families to contain >2, 5, or 10 members. Another requirement could be that the exogenous UIDs of supermutants not be related to any other UID in the library by a one-base difference. Discarding closely related UIDs would eliminate artificial supermutants generated during the library amplification steps (noted in the above paragraph). We routinely discarded such UIDs in our Safe-SeqS analyses, but it made little difference (<1%) in the number of supermutants identified. Specificity for mutations can be further increased by requiring more than one supermutant to identify a position as mutant, as described above for endogenous UIDs. When requiring multiple supermutants, the specificity can be further increased by requiring that each strand of the original double-stranded template contain the mutation or, when libraries are amplified using multiple wells, that rare mutations share an introduced sequence that identifies the well in which the two mutations (i.e., one from each strand) were amplified. Such decisions involve the usual trade-off between specificity and sensitivity. In our experiments with exogenous UIDs (Table 2), we required only one supermutant to identify a position as mutant and included all UID families with more than one member. As the sample preparatory steps required for Safe-SeqS with endogenous UIDs were associated with more error-generating processes than exogenous UIDs, we required two supermutants to identify a position as mutant in the experiments reported in Table 1 and also included all UID families with >1 member.

***Mutation prevalences in normal human tissues.*** The experiments reported in Tables 1 and 2, in which >10,000 templates were assessed, show that mutations are present in the nuclear DNA of normal human cells at a frequency of $3.5 \times 10^{-6}$–$9.0 \times 10^{-6}$ mutants/bp depending on the region analyzed. It is difficult to determine whether this low level represents genuine mutations present in the original templates or the sum of genuine mutations plus artifactual mutations from the error-generating processes described above. Mutation prevalences in human cells have not been widely investigated, in part because they are so infrequent. However, several clever techniques to identify rare mutants have been devised and can in principle be used for comparison. Unfortunately, estimates of human mutation prevalences vary widely, ranging from as many as $10^{-5}$ mutants/bp to as many as $10^{-8}$ mutants/bp (6–12). In several of these studies, the estimates are complicated by the lack of data on the nature of the actual mutations—they could in some cases be caused by losses of whole chromosomes, in others by missense mutations, and in others mainly by nonsense mutations or small insertions or deletions. Additionally, these studies used various sources of normal cells and examined different genes, making direct comparisons difficult. Estimates of the prevalences and rates of mitochondrial DNA mutations similarly vary (13–19). It will be of interest in future work to analyze the same DNA templates and genes with various technologies to determine the basis for these different estimates.

However, let us assume that all of the mutations identified with Safe-SeqS represent genuine mutations present in the original DNA templates from normal cells. What does this tell us about the number of generations through which these cells have proceeded since the organism was conceived? There is a simple relationship between mutation rate and mutation prevalence: The mutation prevalence equals the product of the mutation rate and the number of generations that the cell has gone through since conception. The somatic mutation rate has been determined in previous studies to be $\sim 10^{-9}$ mutants/bp per generation, although this estimate also varies from study to study for reasons related to those mentioned above with respect to mutation prevalence. Combining this literature-derived estimate of mutation rate with our estimates of mutation prevalence suggests that the normal cells analyzed (lymphocytes, lymphoblastoid cell lines, or colonic mucosae) had proceeded through 3,500–8,900 generations, representing cells dividing every 3–7 d for the individuals examined in this study (average age, 65 y).

***Computer simulation of polymerase-introduced errors.*** The timing of mutations introduced by polymerases greatly alters the final number of mutations observed (20). For example, two mutations would differ in prevalence by ~64-fold if introduced 6 cycles apart ($2^6$). Because polymerases introduce mutations in a stochastic manner, a simple Monte Carlo method was used for the simulations. In these simulations, we used the manufacturer's estimate of the Phusion polymerase error rate with an appropriate adjustment for ability of Safe-SeqS to detect mutations in only one strand (4). Note that errors introduced in cycle 19, as well as in the 2 UID assignment cycles, would result in changes in only one strand of the duplex—i.e., result in one supermutant rather than two. In each experiment, we assumed that there was a constant efficiency of amplification given by the total number of templates obtained at the end of the experiment (i.e., if the number of UID families was $N$, then we assumed that the number of templates increased by a factor of $N/2^{19}$ in each cycle). One thousand simulations were performed for each of seven experiments, and the results are reported in Table S2.


1. Herman DS, et al. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* 6:507–510.
2. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
3. Chory J, Pollard JD Jr. (2001) Separation of small DNA fragments by conventional gel electrophoresis. *Curr Protoc Mol Biol* Chap 2:Unit2.7.
4. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* 112:29–35.
5. Collins AR (1999) Oxidative DNA damage, antioxidants, and cancer. *BioEssays* 21:238–246.
6. Morley AA, Cox S, Holliday R (1982) Human lymphocytes resistant to 6-thioguanine increase with age. *Mech Ageing Dev* 19:21–26.
7. Trainor KJ, et al. (1984) Mutation frequency in human lymphocytes increases with age. *Mech Ageing Dev* 27:83–86.
8. Grist SA, McCarron M, Kutlaca A, Turner DR, Morley AA (1992) In vivo human somatic mutation: Frequency and spectrum with age. *Mutat Res* 266:189–196.
9. Williams GT, Geraghty JM, Campbell F, Appleton MA, Williams ED (1995) Normal colonic mucosa in hereditary non-polyposis colorectal cancer shows no generalised increase in somatic mutation. *Br J Cancer* 71:1077–1080.
10. Campbell F, Appleton MA, Shields CJ, Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right- and left-sided sporadic colorectal carcinomas. *J Pathol* 186:31–35.
11. Araten DJ, Nafa K, Pakdeesuwan K, Luzzatto L (1999) Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals. *Proc Natl Acad Sci USA* 96: 5209–5214.
12. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.
13. Monnat RJ, Jr., Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. *Proc Natl Acad Sci USA* 82:2895–2899.
14. Bodenteich A, Mitchell LG, Merril CR (1991) A lifetime of retinal light exposure does not appear to increase mitochondrial mutations. *Gene* 108:305–309.
15. Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet* 59:501–509.
16. Khrapko K, et al. (1997) Mitochondrial mutational spectra in human cells and tissues. *Proc Natl Acad Sci USA* 94:13798–13803.
17. Heyer E, et al. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* 69:1113–1126.
18. Howell N, et al. (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. *Am J Hum Genet* 72:659–670.

19. Taylor RW, et al. (2003) Mitochondrial DNA mutations in human colonic crypt stem cells. *J Clin Invest* 112:1351–1360.

20. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.

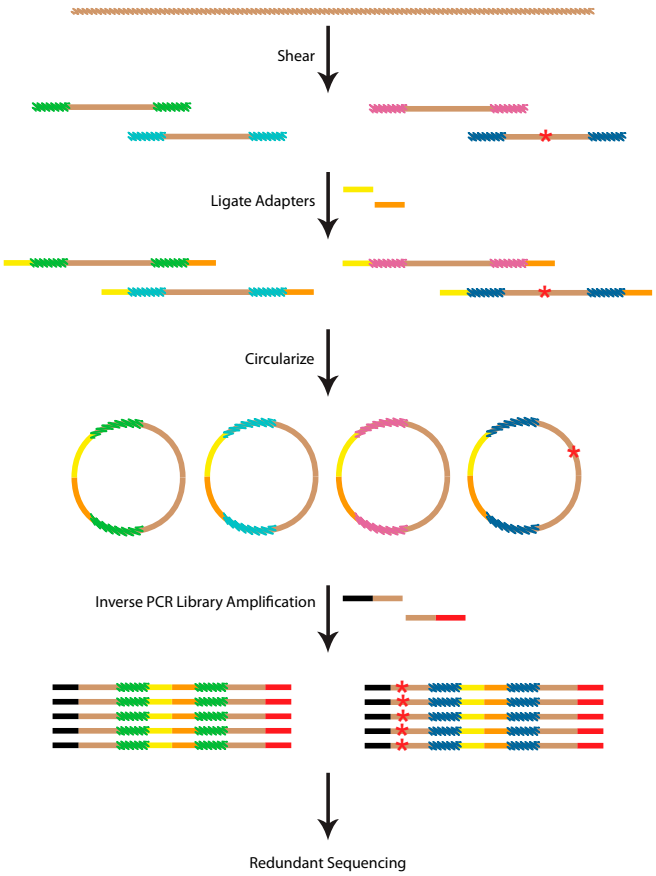


**Fig. S1.** Safe-SeqS with endogenous UIDs plus inverse PCR. The sequences of the ends of fragments produced by random shearing serve as unique identifiers (UIDs) (variously colored bars). These fragments are ligated to adapters (yellow and orange bars) as in a standard Illumina library preparation. One uniquely tagged fragment is produced from each strand of the double-stranded template; only one strand is shown. Following circularization with a ligase, inverse PCR is performed with gene-specific primers that also contain 5′ "grafting" sequences (black and red bars). This PCR produces UID families that are directly sequenced. Supermutants are defined as in Fig. 1.

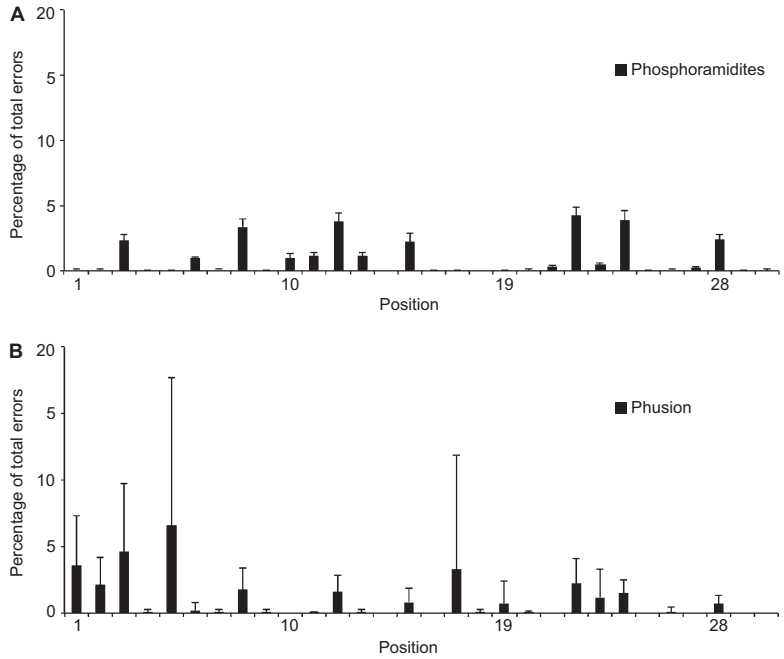


**Fig. S2.** Single-base substitutions position vs. error frequency in oligonucleotides synthesized with phosphoramidites and Phusion. A representative portion of the same 31-base DNA fragment synthesized with phosphoramidites (*A*) or Phusion polymerase (*B*) was analyzed by Safe-SeqS. The means and SDs for seven independent experiments of each type are plotted. Averages of 1,721 ± 383 and 196 ± 143 single-base substitution supermutants were identified in the phosphoramidite-synthesized and Phusion-generated fragments, respectively. The *y* axis indicates the fraction of the total errors at the indicated position. Note that the errors in the phosphoramidite-synthesized DNA fragment were consistent among the seven replicates, as would be expected if the errors were systematically introduced during the synthesis itself. In contrast, the errors in the Phusion-generated fragments appeared to be heterogeneous among samples, as expected from a stochastic process (20).

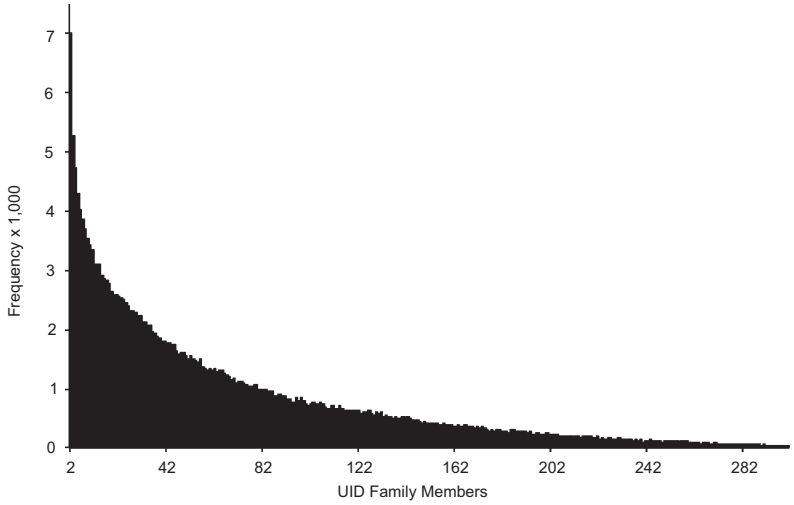


**Fig. S3.** UID family member distribution. The exogenous UID strategy depicted in Fig. 3 was used to produce PCR fragments from a region of *CTNNB1* from three normal, unrelated individuals (Table 2, *CTNNB1* mutations in DNA from normal human cells); a representative example of the UID families with ≤300 members (99% of total UID families) generated from one individual is shown. The *y* axis indicates the number of different UID families that contained the number of family members shown on the *x* axis.

**Table S1. Fraction of single-base substitutions, insertions, and deletions with exogenous UIDs**

| | Mean | SD |
|---|---|---|
| Polymerase fidelity | | |
| Conventional analysis of seven replicates | | |
| Total mutations identified | 198,638 | 22,515 |
| % mutations represented by single-base substitutions | 98.90 | 0.16 |
| % mutations represented by deletions | 1.05 | 0.16 |
| % mutations represented by insertions | 0.05 | 0.01 |
| Safe-SeqS analysis of seven replicates | | |
| Total supermutants identified | 197 | 143 |
| % supermutants represented by single-base substitutions | 99.10 | 2.06 |
| % supermutants represented by deletions | 0.90 | 2.06 |
| % supermutants represented by insertions | 0.00 | 0.00 |
| *CTNNB1* mutations in DNA from normal human cells | | |
| Conventional analysis of three individuals | | |
| Total mutations identified | 118,488 | 11,357 |
| % mutations represented by single-base substitutions | 97.38 | 0.30 |
| % mutations represented by deletions | 2.62 | 0.30 |
| % mutations represented by insertions | 0.00 | 0.00 |
| Safe-SeqS analysis of three individuals | | |
| Total supermutants identified | 99 | 78 |
| % supermutants represented by single-base substitutions | 99.52 | 0.84 |
| % supermutants represented by deletions | 0.48 | 0.84 |
| % supermutants represented by insertions | 0.00 | 0.00 |
| Mitochondrial mutations in DNA from normal human cells | | |
| Conventional analysis of seven individuals | | |
| Total mutations identified | 30,599 | 12,970 |
| % mutations represented by single-base substitutions | 98.42 | 0.64 |
| % mutations represented by deletions | 1.58 | 0.64 |
| % mutations represented by insertions | 0.00 | 0.00 |
| Safe-SeqS analysis of seven individuals | | |
| Total supermutants identified | 135 | 61 |
| % supermutants represented by single-base substitutions | 98.66 | 1.39 |
| % supermutants represented by deletions | 1.34 | 1.39 |
| % supermutants represented by insertions | 0.00 | 0.00 |

**Table S2. Observed and expected number of errors generated by Phusion polymerase**

| | Observed | Expected (mean ± SD)* |
|---|---|---|
| Experiment 1 | | |
| Mutations represented by 1 supermutant | 10 | 19 ± 3.7 |
| Mutations represented by 2 supermutants | 8 | 5.8 ± 2.3 |
| Mutations represented by 3 supermutants | 4 | 1.3 ± 1.1 |
| Mutations represented by 4 supermutants | 4 | 1.8 ± 1.3 |
| Mutations represented by 5 supermutants | 2 | 0.61 ± 0.75 |
| Mutations represented by 6 supermutants | 2 | 0.22 ± 0.44 |
| Mutations represented by 7 supermutants | 0 | 0.01 ± 0.10 |
| Mutations represented by 8 supermutants | 0 | 0.87 ± 0.86 |
| Mutations represented by 9 supermutants | 2 | 0.28 ± 0.51 |
| Mutations represented by 10 supermutants | 0 | 0.14 ± 0.38 |
| Mutations represented by >10 supermutants | 3 | 1.5 ± 2.7 |
| Distinct mutations | 35 | 32 ± 4.2 |
| Experiment 2 | | |
| Mutations represented by 1 supermutant | 19 | 23 ± 4.1 |
| Mutations represented by 2 supermutants | 5 | 9.5 ± 2.8 |
| Mutations represented by 3 supermutants | 4 | 2.7 ± 1.6 |
| Mutations represented by 4 supermutants | 7 | 2.7 ± 1.7 |
| Mutations represented by 5 supermutants | 2 | 0.88 ± 0.94 |
| Mutations represented by 6 supermutants | 1 | 0.40 ± 0.60 |
| Mutations represented by 7 supermutants | 3 | 0.16 ± 0.42 |
| Mutations represented by 8 supermutants | 1 | 0.99 ± 1.0 |
| Mutations represented by 9 supermutants | 1 | 0.39 ± 0.68 |
| Mutations represented by 10 supermutants | 0 | 0.17 ± 0.43 |
| Mutations represented by >10 supermutants | 9 | 1.8 ± 3.4 |
| Distinct mutations | 52 | 43 ± 5.1 |
| Experiment 3 | | |
| Mutations represented by 1 supermutant | 7 | 17 ± 3.4 |
| Mutations represented by 2 supermutants | 9 | 5.4 ± 2.0 |
| Mutations represented by 3 supermutants | 4 | 1.2 ± 1.1 |
| Mutations represented by 4 supermutants | 4 | 1.7 ± 1.4 |
| Mutations represented by 5 supermutants | 2 | 0.50 ± 0.70 |
| Mutations represented by 6 supermutants | 0 | 0.17 ± 0.45 |
| Mutations represented by 7 supermutants | 1 | 0.03 ± 0.17 |
| Mutations represented by 8 supermutants | 0 | 0.59 ± 0.74 |
| Mutations represented by 9 supermutants | 0 | 0.24 ± 0.50 |
| Mutations represented by 10 supermutants | 1 | 0.07 ± 0.29 |
| Mutations represented by >10 supermutants | 5 | 1.5 ± 2.6 |
| Distinct mutations | 33 | 28 ± 3.7 |
| Experiment 4 | | |
| Mutations represented by 1 supermutant | 7 | 15 ± 3.7 |
| Mutations represented by 2 supermutants | 8 | 4.1 ± 1.7 |
| Mutations represented by 3 supermutants | 2 | 0.70 ± 0.74 |
| Mutations represented by 4 supermutants | 1 | 1.5 ± 1.3 |
| Mutations represented by 5 supermutants | 3 | 0.21 ± 0.52 |
| Mutations represented by 6 supermutants | 2 | 0.08 ± 0.27 |
| Mutations represented by 7 supermutants | 1 | 0.0 ± 0.0 |
| Mutations represented by 8 supermutants | 2 | 0.65 ± 0.77 |
| Mutations represented by 9 supermutants | 2 | 0.17 ± 0.43 |
| Mutations represented by 10 supermutants | 0 | 0.05 ± 0.22 |
| Mutations represented by >10 supermutants | 1 | 0.92 ± 2.1 |
| Distinct mutations | 29 | 23 ± 3.2 |
| Experiment 5 | | |
| Mutations represented by 1 supermutant | 9 | 23 ± 4.1 |
| Mutations represented by 2 supermutants | 6 | 9.5 ± 2.8 |
| Mutations represented by 3 supermutants | 5 | 2.7 ± 1.6 |
| Mutations represented by 4 supermutants | 3 | 2.7 ± 1.7 |
| Mutations represented by 5 supermutants | 6 | 0.88 ± 0.94 |
| Mutations represented by 6 supermutants | 2 | 0.40 ± 0.60 |
| Mutations represented by 7 supermutants | 1 | 0.16 ± 0.42 |
| Mutations represented by 8 supermutants | 2 | 0.99 ± 1.0 |
| Mutations represented by 9 supermutants | 2 | 0.39 ± 0.68 |
| Mutations represented by 10 supermutants | 3 | 0.17 ± 0.43 |

**Table S2. Cont.**

| | Observed | Expected (mean ± SD)* |
|---|---|---|
| Mutations represented by >10 supermutants | 7 | 1.8 ± 3.4 |
| Distinct mutations | 46 | 43 ± 5.1 |
| Experiment 6 | | |
| Mutations represented by 1 supermutant | 4 | 6.7 ± 2.8 |
| Mutations represented by 2 supermutants | 7 | 1.5 ± 1.2 |
| Mutations represented by 3 supermutants | 1 | 0.10 ± 0.33 |
| Mutations represented by 4 supermutants | 2 | 0.60 ± 0.82 |
| Mutations represented by 5 supermutants | 0 | 0.07 ± 0.26 |
| Mutations represented by 6 supermutants | 0 | 0.01 ± 0.10 |
| Mutations represented by 7 supermutants | 1 | 0.0 ± 0.0 |
| Mutations represented by 8 supermutants | 1 | 0.39 ± 0.60 |
| Mutations represented by 9 supermutants | 0 | 0.01 ± 0.10 |
| Mutations represented by 10 supermutants | 0 | 0.0 ± 0.0 |
| Mutations represented by >10 supermutants | 2 | 0.50 ± 1.1 |
| Distinct mutations | 18 | 9.9 ± 1.4 |
| Experiment 7 | | |
| Mutations represented by 1 supermutant | 8 | 2.9 ± 1.6 |
| Mutations represented by 2 supermutants | 2 | 0.61 ± 0.79 |
| Mutations represented by 3 supermutants | 0 | 0.04 ± 0.24 |
| Mutations represented by 4 supermutants | 0 | 0.41 ± 0.59 |
| Mutations represented by 5 supermutants | 1 | 0.01 ± 0.10 |
| Mutations represented by 6 supermutants | 0 | 0.0 ± 0.0 |
| Mutations represented by 7 supermutants | 0 | 0.0 ± 0.0 |
| Mutations represented by 8 supermutants | 0 | 0.14 ± 0.35 |
| Mutations represented by 9 supermutants | 0 | 0.01 ± 0.10 |
| Mutations represented by 10 supermutants | 0 | 0.0 ± 0.0 |
| Mutations represented by >10 supermutants | 0 | 0.32 ± 0.93 |
| Distinct mutations | 11 | 4.5 ± 0.62 |

*See *SI Materials and Methods* for details of the simulations.

**Table S3. Phosphoramidite- vs. Phusion-synthesized DNA: transitions vs. transversions comparison**

| | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 5 | Exp. 6 | Exp. 7 | Average | SD |
|---|---|---|---|---|---|---|---|---|---|
| Phosphoramidites | | | | | | | | | |
| Transition supermutants | 496 | 509 | 471 | 396 | 323 | 273 | 470 | 420 | 92 |
| Transversion supermutants | 1,494 | 1,499 | 1,521 | 1,154 | 944 | 907 | 1,626 | 1,306 | 298 |
| *P* value* | 3.4*E*-05 | | | | | | | | |
| Phusion | | | | | | | | | |
| Transition supermutants | 63 | 275 | 127 | 5 | 87 | 182 | 103 | 120 | 87 |
| Transversion supermutants | 14 | 124 | 77 | 12 | 57 | 191 | 63 | 77 | 63 |
| *P* value* | 0.08 | | | | | | | | |

**P* values were calculated using a two-tailed paired *t* test. Exp., experiment.

**Table S4. Oligonucleotides used in this study**

| | Sequence |
|---|---|
| **Endogenous UIDs** | |
| Capture | |
| Adapter, strand 1 | /5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG |
| Adapter, strand 2 | ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for | AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, rev | CAAGCAGAAGACGGCATACGAGATCTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T |
| Postcapture amplification, for | AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Postcapture amplification, rev | CAAGCAGAAGACGGCATACGAGATCTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T |
| Sequencing primer, read 1 (Illumina) | ACACTCTTTCCCTACACGACGCTCTTCCGATCT |
| Sequencing primer, read 2 (Illumina) | CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT |
| Inverse PCR | |
| Adapter, strand 1 | /5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG |
| Adapter, strand 2 | ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-1 | /5Phos/CGTGATACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-2 | /5Phos/ACATCGACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-3 | /5Phos/GCCTAAACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-4 | /5Phos/TGGTCAACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-5 | /5Phos/CACTGTACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-6 | /5Phos/ATTGGCACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-7 | /5Phos/GATCTGACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-8 | /5Phos/TCAAGTACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-9 | /5Phos/CTGATCACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-10 | /5Phos/AAGCTAACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-11 | /5Phos/GTAGCCACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, for-12 | /5Phos/TACAAGACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T |
| Whole-genome amplification, rev | /5Phos/CTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T |
| Inverse PCR, antisense | AATGATACGGCGACCACCGAGATCTACACCAGCAGGCCTTATAATAAAAATAATGA |
| Inverse PCR, for | CAAGCAGAAGACGGCATACGAGATTGACTGAATATAAACTTGTGGTAGTTG |
| Sequencing primer 1 (to read internal sequences) | ACACTCTTTCCCTACACGACGCTCTTCCGATCT |
| Sequencing primer 2 (to read internal sequences) | CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT |
| Index primer 1 (to read experiment indexes) | CGGAAGAGCGTCGTGTAGGGAAAGAGTGT |
| Index primer 2 (to read experiment indexes) | CGGAAGAGCGGTTCAGCAGGAATGCCGAG |
| **Exogenous UIDs** | |
| Polymerase fidelity | |
| Digital PCR amplification, for | GGTTACAGGCTCATGATGTAACC |
| Digital PCR amplification, rev | GATACCAGCTTGGTAATGGCA |
| UID assignment amplification, for | CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNNGGTTACAGGCTCATGATGTAACC |
| UID assignment amplification, rev | CACACAGGAAACAGCTATGACCATGGATACCAGCTTGGTAATGGCA |
| Library amplification, for-1 | AATGATACGGCGACCACCGAGATCTACACCGTGATCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-2 | AATGATACGGCGACCACCGAGATCTACACACATCGCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-3 | AATGATACGGCGACCACCGAGATCTACACGCCTAACGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-4 | AATGATACGGCGACCACCGAGATCTACACTGGTCACGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-5 | AATGATACGGCGACCACCGAGATCTACACCACTGTCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-6 | AATGATACGGCGACCACCGAGATCTACACATTGGCCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-7 | AATGATACGGCGACCACCGAGATCTACACGATCTGCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-8 | AATGATACGGCGACCACCGAGATCTACACTCAAGTCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-9 | AATGATACGGCGACCACCGAGATCTACACCTGATCCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-10 | AATGATACGGCGACCACCGAGATCTACACAAGCTACGACGTAAAACGACGGCCA*G*T |
| Library amplification, rev | CAAGCAGAAGACGGCATACGAGATCACACAGGAAACAGCTATGACCA*T*G |
| Sequencing primer (to read UID and internal sequences) | CGACGTAAAACGACGGCCAGT |
| Index primer (to read experiment indexes) | ACTGGCCGTCGTTTTACGTCG |

**Table S4. Cont.**

| | Sequence |
|---|---|
| *CTNNB1* mutations in DNA from normal human cells | |
| UID assignment amplification, for | CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNNNNGCAGCAACAGTCTTACCTGGACT |
| UID assignment amplification, rev | CACACAGGAAACAGCTATGACCATGTCCACATCCTCTTCCTCAGGATT |
| Library amplification, for | AATGATACGGCGACCACCGAGATCTACACCGACGTAAAACGACGGCCA*G*T |
| Library amplification, rev-1 | CAAGCAGAAGACGGCATACGAGATATCACGCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-2 | CAAGCAGAAGACGGCATACGAGATCGATGTCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-3 | CAAGCAGAAGACGGCATACGAGATTGACCACACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-4 | CAAGCAGAAGACGGCATACGAGATGCCAATCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-5 | CAAGCAGAAGACGGCATACGAGATCAGATCCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-6 | CAAGCAGAAGACGGCATACGAGATACTTGACACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-7 | CAAGCAGAAGACGGCATACGAGATGATCAGCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-8 | CAAGCAGAAGACGGCATACGAGATTAGCTTCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-9 | CAAGCAGAAGACGGCATACGAGATGGCTACCACACAGGAAACAGCTATGACCA*T*G |
| Library amplification, rev-10 | CAAGCAGAAGACGGCATACGAGATCTTGTACACACAGGAAACAGCTATGACCA*T*G |
| Sequencing primer (to read UID and internal sequences) | CGACGTAAAACGACGGCCAGT |
| Index primer (to read experiment indexes) | CATGGTCATAGCTGTTTCCTGTGTG |
| Mitochondrial mutations in DNA from normal human cells | |
| UID assignment amplification, for | CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNNNNTTACCGAGAAAGCTCACAAGAA |
| UID assignment amplification, rev | CACACAGGAAACAGCTATGACCATGATGCTAAGGCGAGGATGAAA |
| Library amplification, for-1 | AATGATACGGCGACCACCGAGATCTACACACATCGCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-2 | AATGATACGGCGACCACCGAGATCTACACGCCTAACGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-3 | AATGATACGGCGACCACCGAGATCTACACTGGTCACGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-4 | AATGATACGGCGACCACCGAGATCTACACATTGGCCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-5 | AATGATACGGCGACCACCGAGATCTACACGATCTGCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-6 | AATGATACGGCGACCACCGAGATCTACACTCAAGTCGACGTAAAACGACGGCCA*G*T |
| Library amplification, for-7 | AATGATACGGCGACCACCGAGATCTACACCTGATCCGACGTAAAACGACGGCCA*G*T |
| Library amplification, rev | CAAGCAGAAGACGGCATACGAGATCACACAGGAAACAGCTATGACCA*T*G |
| Sequencing primer 1 (to read UIDs) | CGACGTAAAACGACGGCCAGT |
| Sequencing primer 2 (to read internal sequences) | CCTAATTCCCCCCATCCTTAC |
| Index primer (to read experiment indexes) | ACTGGCCGTCGTTTTACGTCG |
| Analysis of phosphoramidite oligonucleotide composition | |
| Synthesized template, WT | GGTTACAGGCTCATGATGTAACCTCTGTGTCTTGGTG**T**AACTTTAAAACATATTTTTGCCATTACCAAGCTGGTATC |
| Synthesized template, mut (*S* = 50/50 mix of C and G) | GGTTACAGGCTCATGATGTAACCTCTGTGTCTTGGTG**S**AACTTTAAAACATATTTTTGCCATTACCAAGCTGGTATC |
| UID assignment amplification, for | ACACTCTTTCCCTACACGACGCTCNNNNNNNNNNNNNGGTGAGTCTGTGCAGGCAT |
| UID assignment amplification, rev | CTCGAGCACTGTCCTGACTGAGACGATACCAGCTTGGTAATGGCA |
| Library amplification, for | AATGATACGGCGACCACCGAGATCTACACCGTGATACACTCTTTCCCTACACGACGC*T*C |
| Library amplification, rev | CAAGCAGAAGACGGCATACGAGATCTCGAGCACTGTCCTGACTGAG*A*C |
| Sequencing primer (to read UID and internal sequences) | ACACTCTTTCCCTACACGACGCTC |

Colors: blue, region complementary to templates; green, template-specific UID sequence; black, universal sequence; purple, experiment-specific index sequence; red, Illumina grafting primers (for hybridization to flow cell). /5Phos/, 5′ phosphate; *, phosphorothioate linkage; for, forward; rev, reverse.