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.

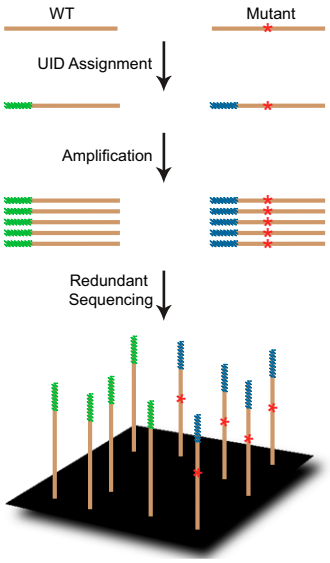

Fig. 1. Essential elements of Safe-SeqS. In the first step, each fragment to be analyzed is assigned a unique identification (UID) DNA sequence (green or blue bars). In the second step, the uniquely tagged fragments are amplified, producing UID families, each member of which has the same UID. A supermutant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

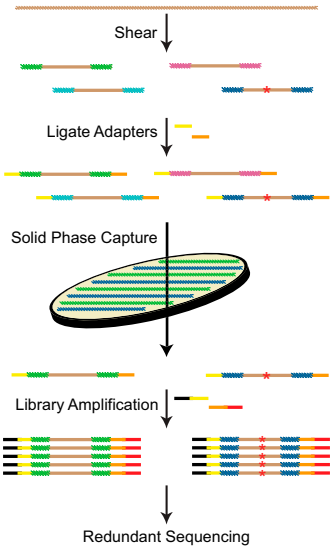

Fig. 2. Safe-SeqS with endogenous UIDs plus capture. The sequences of the ends of each fragment produced by random shearing (variously colored bars) serve as the unique identifiers (UIDs). These fragments are ligated to adapters (yellow and orange bars) so they can subsequently be amplified by PCR. One 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 2.4×10^{-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 (2.4×10^{-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 3.5×10^{-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 2.3×10^{-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

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.

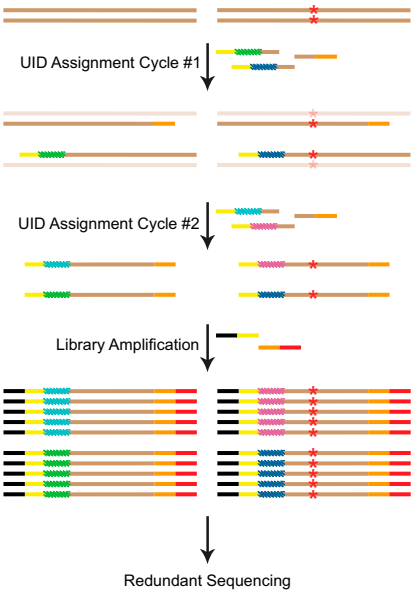

Fig. 3. Safe-SeqS with exogenous UIDs. DNA (sheared or unsheared) is amplified with a set of gene-specific primers. One of the primers has a random DNA sequence (e.g., a set of 14 Ns) that forms the unique identifier (UID) (variously colored bars), located 5′ to its gene-specific sequence, and both have 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 4.4×10^{-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 9.1×10^{-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 4.5×10^{-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 ± 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 ± 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-

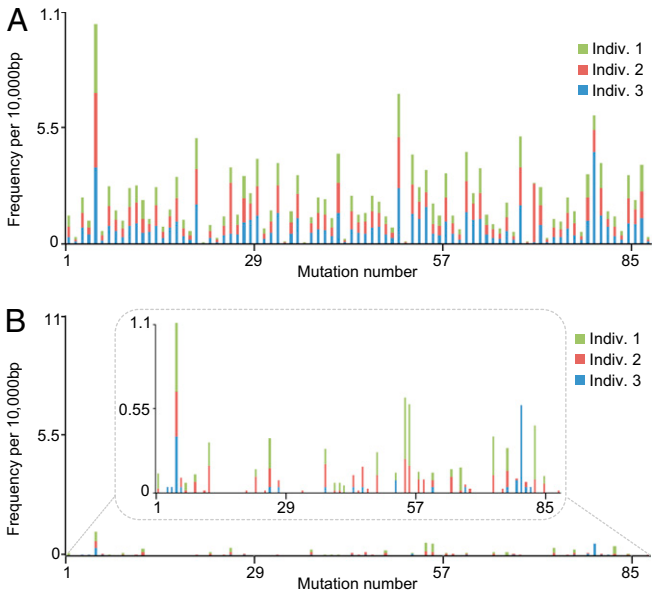

Fig. 4. Single-base substitutions identified by conventional and Safe-SeqS analysis. The exogenous UID strategy depicted in Fig. 3 was used to produce PCR fragments from the *CTNNB1* gene of three normal, unrelated individuals. Mutation numbers represent one of 87 possible single-base substitutions (3 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 2.0×10^{-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: 2.0×10^{-4}–2.4×10^{-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 1.4×10^{-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., 8.8×10^{-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.