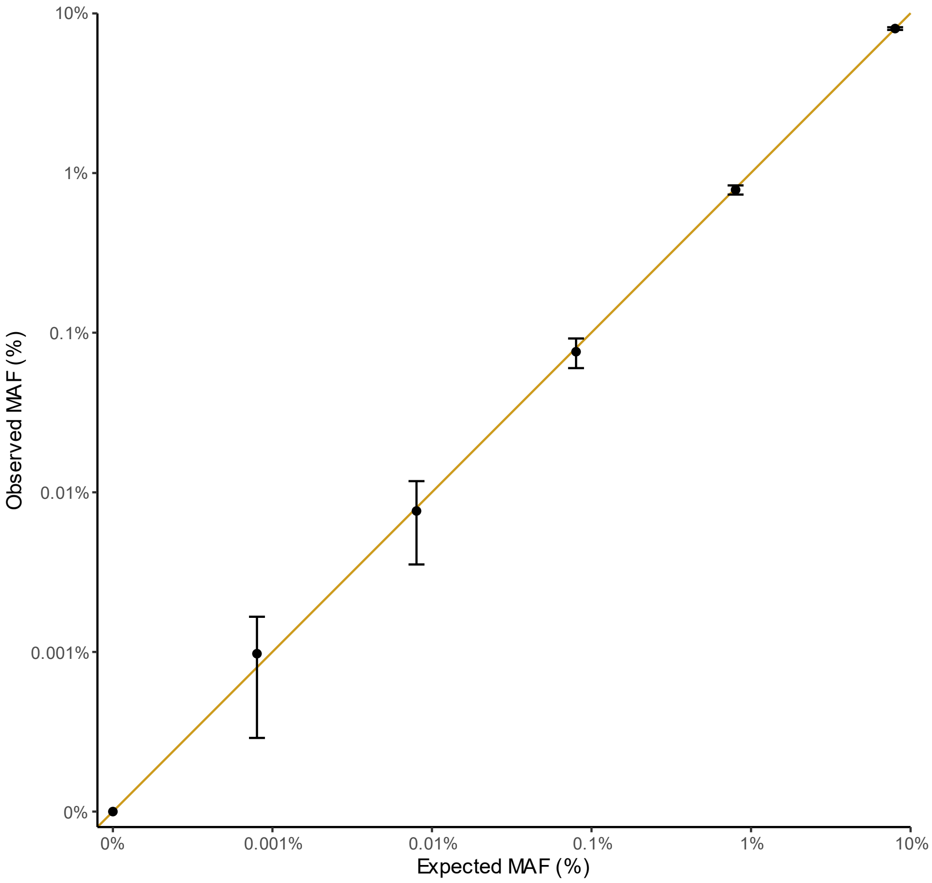
nature portfolio

https://doi.org/10.1038/s41587-021-00900-z

## Supplementary information

# Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands

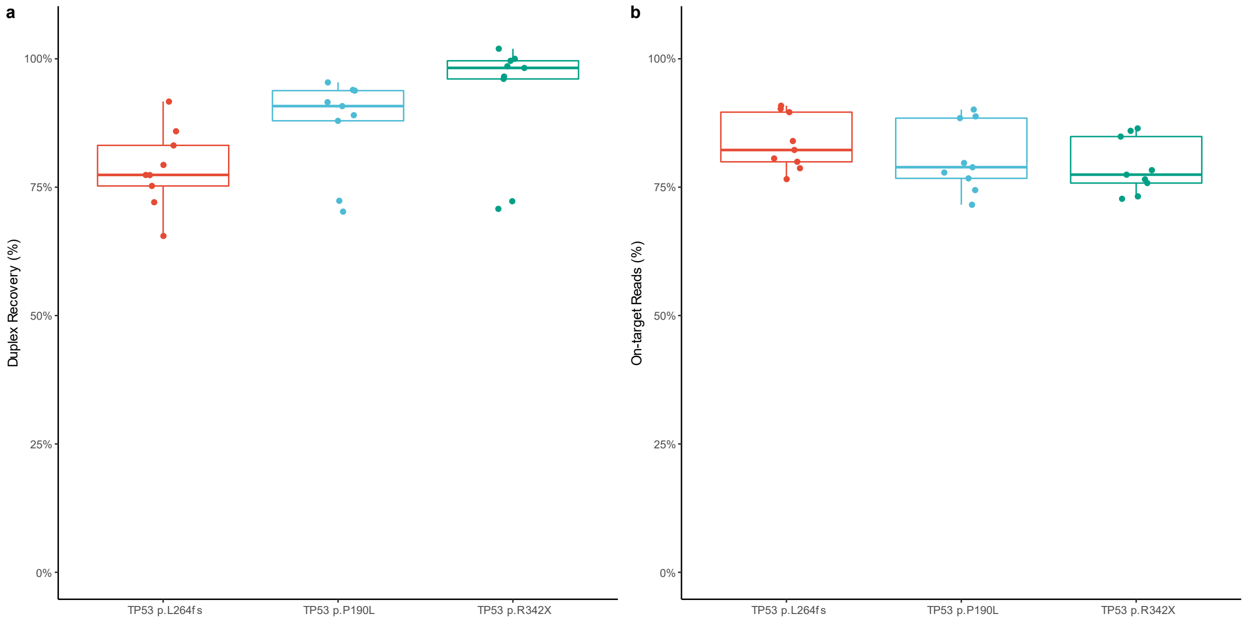
In the format provided by the authors and unedited



**Supplementary Figure 1**

**Analytical performance of SaferSeqS.**

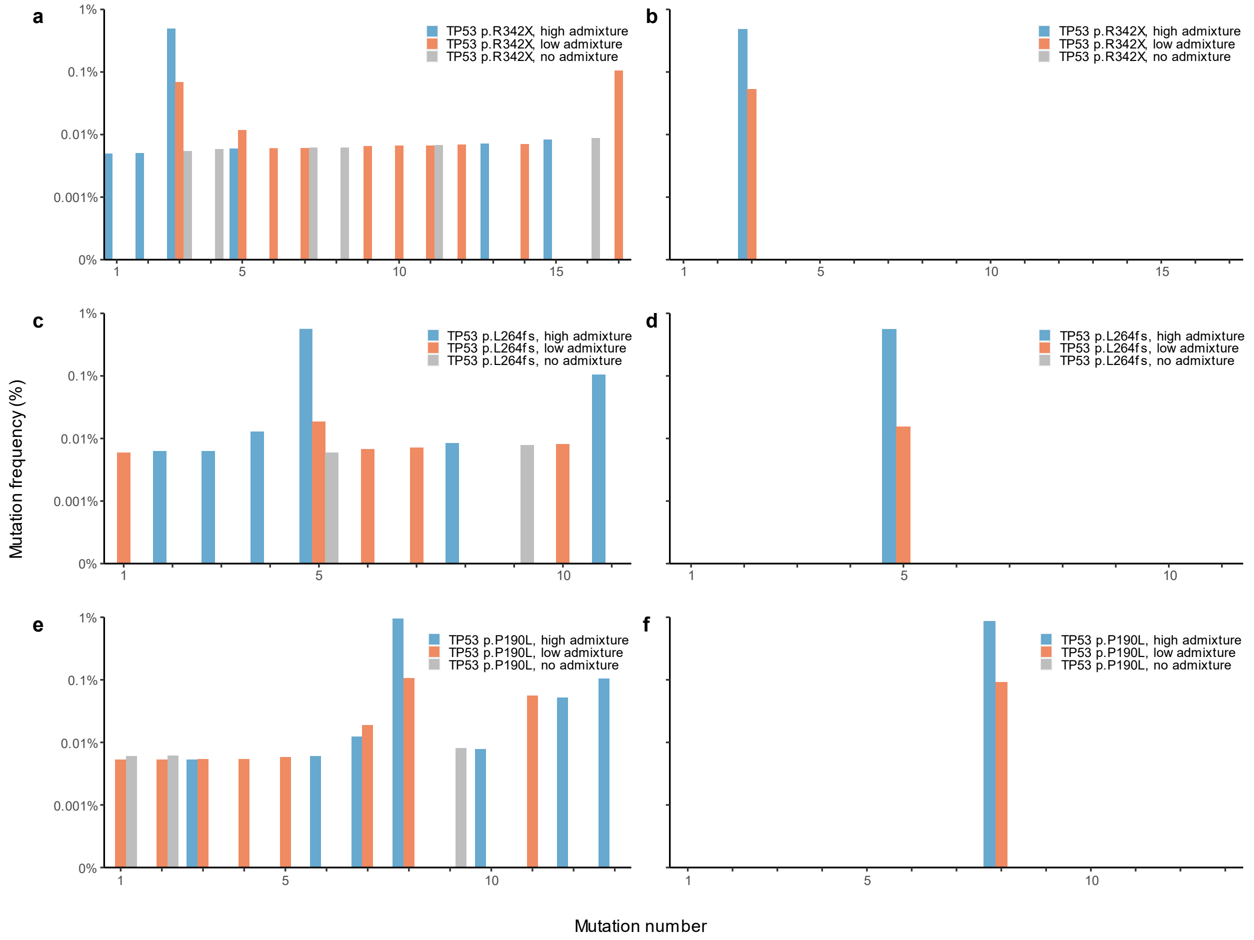
Mutant allele frequencies (MAF) determined by SaferSeqS versus the expected frequencies when DNA from a cancer containing a known mutation was mixed with leukocyte DNA from a healthy donor at ratios varying from 8% to 0.0008%. A 0% control sample was also assayed to determine specificity for the mutation of interest. The solid line represents a fit of a linear regression model in which the y-intercept was fixed at zero (slope = 1.004, adjusted $R^2 > 0.999$, $P = 2.42 \times 10^{-14}$, two-sided F-test). Data are presented as mean allele frequencies ± s.e.m.



## Supplementary Figure 2

### High duplex recovery and efficient target enrichment with SaferSeqS.

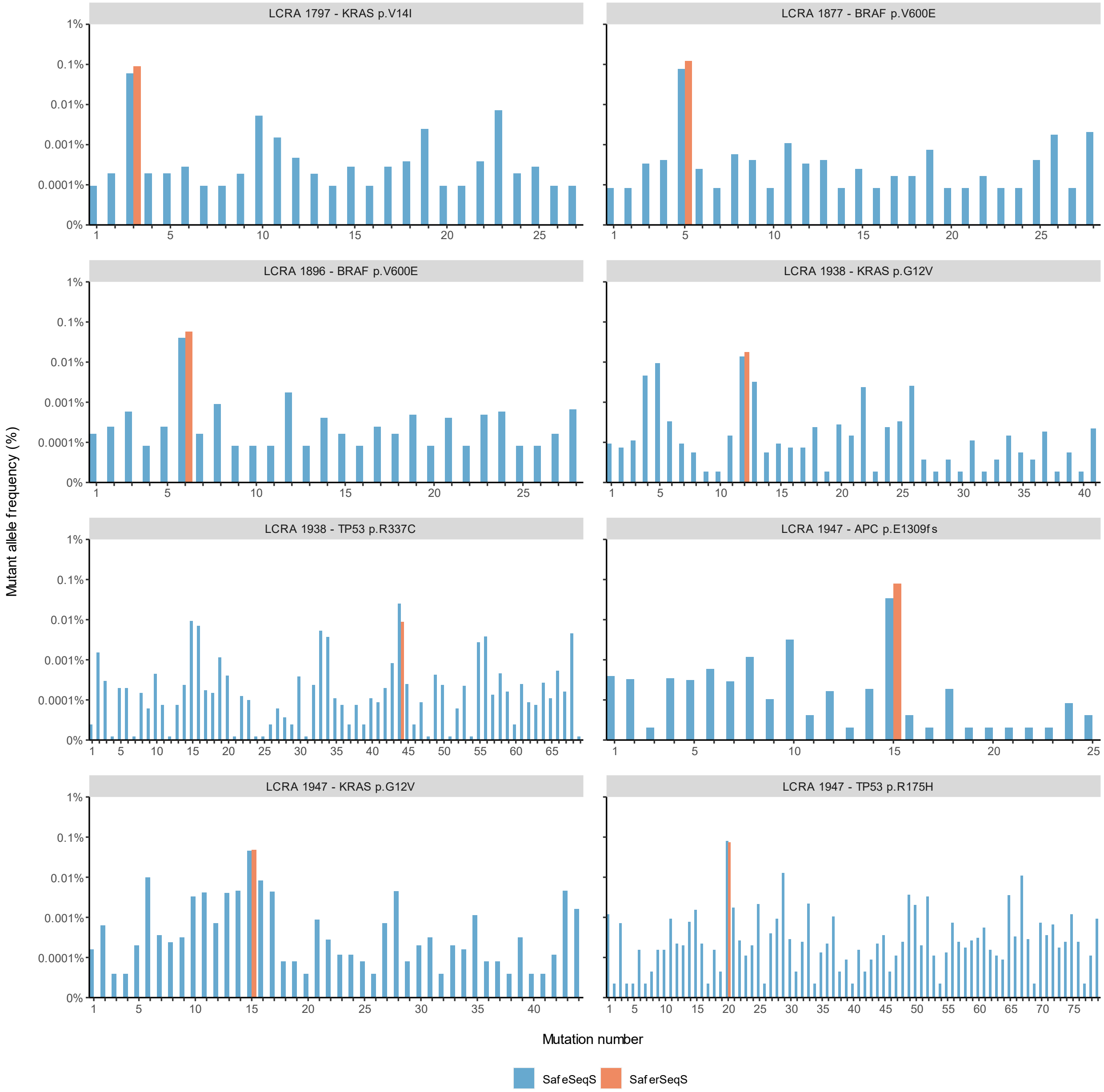
Thirty-three ng of admixed cfDNA samples were assayed for one of three different mutations in *TP53* (p.L264fs, p.P190L, or p.R342X). Three libraries for each of the three dilutions were prepared per cfDNA admixture, each containing ~11 ng of cfDNA (n = 9 libraries/admixture). **(a)** The median number of duplex families (i.e., both Watson and Crick strands containing the same endogenous and exogenous barcodes) was 89% (range: 65% to 102%) of the number of original template molecules. **(b)** The median fraction of on-target reads was 80% (range: 72% to 91%). Lower and upper hinges correspond to the 25th and 75th percentile, whiskers extend to 1.5 times the interquartile range. Individual data points are overlaid with random scatter.



## Supplementary Figure 3

**Errors in SaferSeqS as compared to those of strand-agnostic, ligation-based molecular barcoding methods.**

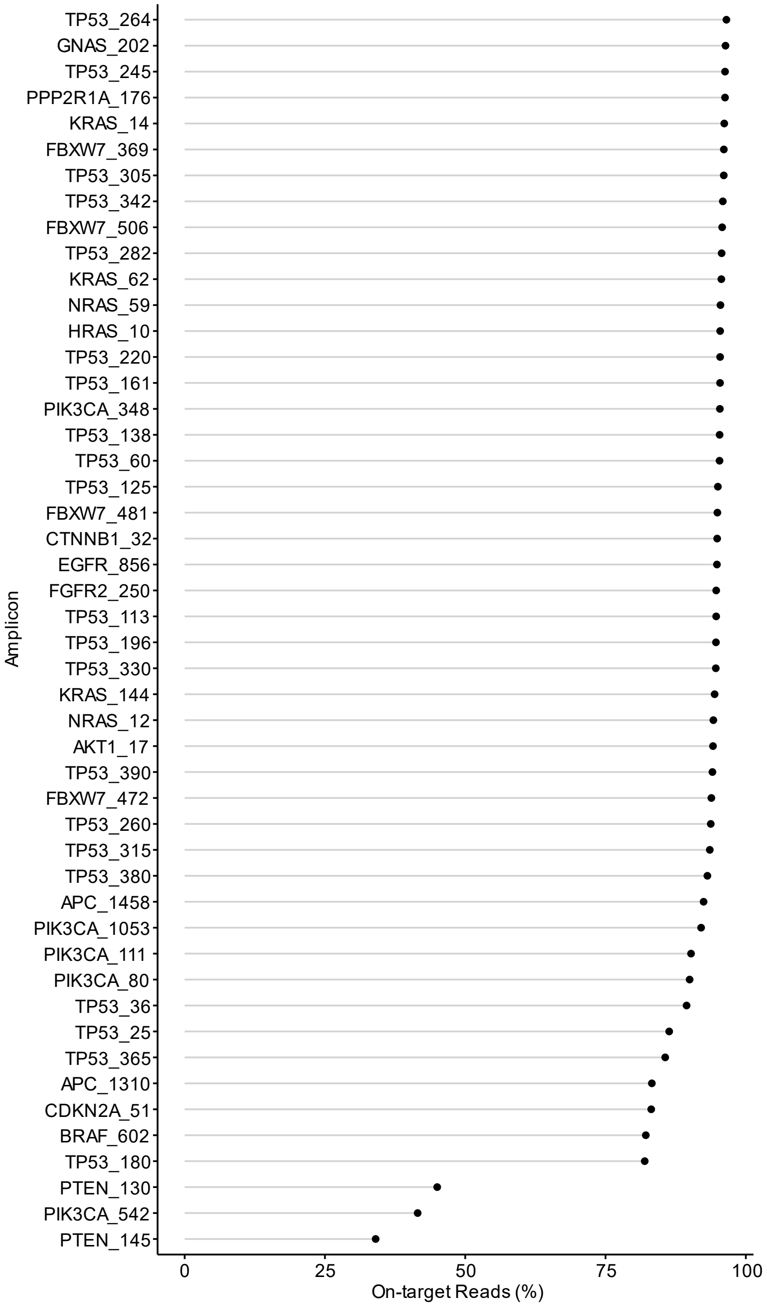
Analysis of 33 ng of plasma cell-free DNA from healthy individuals admixed with cell-free plasma DNA from a cancer patient. The mixtures were created to generate a high frequency (~0.5-1%) of mutation (blue bars), low frequency (~0.01-0.1%) of mutation (orange bars), or no mutation (grey bars). The admixed *TP53* p.R342X sample was assayed with SaferSeqS but **(a)** strand information was ignored in the analysis to mimic strand-agnostic, ligation-based molecular barcoding methods or **(b)** strand information was considered during mutation calling. Similarly, the admixed *TP53* p.L264fs sample was assayed **(c)** without consideration of strand information and **(d)** with SaferSeqS. The admixed *TP53* p.P190L sample was similarly assayed **(e)** without consideration of strand information and **(f)** with SaferSeqS. Mutations with a depth of >100 UIDs are shown; mutation numbers are defined in **Supplementary Table 3**.



## Supplementary Figure 4

## Evaluation of plasma samples from cancer patients.

Plasma cell-free DNA samples from five cancer patients harboring eight known mutations at frequencies between 0.01% and 0.1% were assayed with a previously described, PCR-based molecular barcoding method ("SafeSeqS" rather than "SaferSeqS", blue bars) and with SaferSeqS (orange bars).  Mutation numbers are defined in **Supplementary Table 4**.



## Supplementary Figure 5

**Performance of the 48 primer pairs used in a multiplex panel to assay regions of driver genes commonly mutated in cancer.**

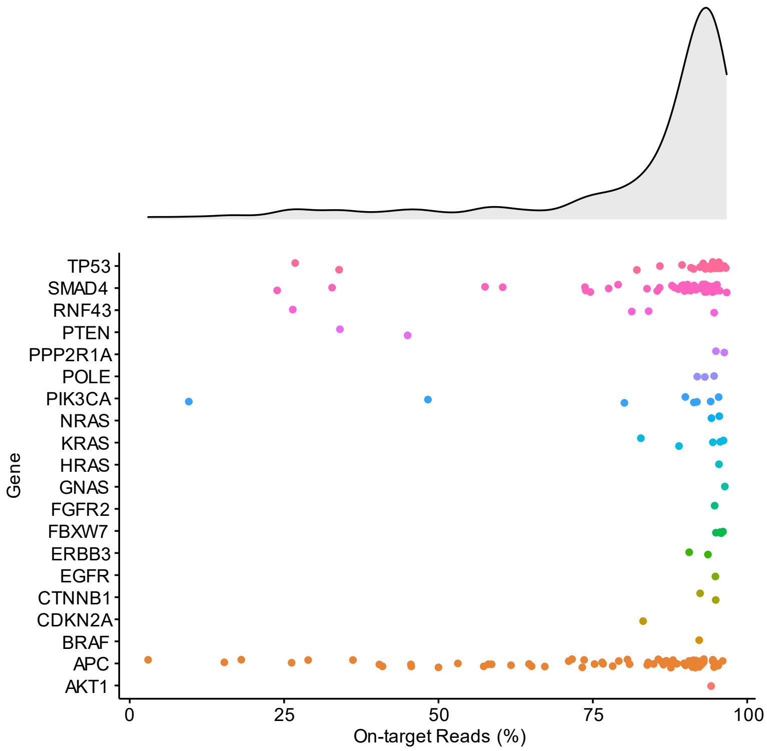
The proportion of on-target reads (i.e. the fraction of total reads that map to the intended target) for each of the 48 SaferSeqS primer pairs used in the strand-specific PCRs.  Primers were used at the concentrations denoted in **Supplementary Table 5** in each gene-specific PCR (**Methods**).

|  |  | Multiplex assay |  |
|---|---|---|---|
|  |  | Detectable | Not Detectable |
| Singleplex assay | Detectable | 42 | 1 |
|  | Not Detectable | 1 | 31 |

**Supplementary Figure 6**

**Confusion matrix of patient-level sensitivities achieved with multiplex and single amplicon assays.**

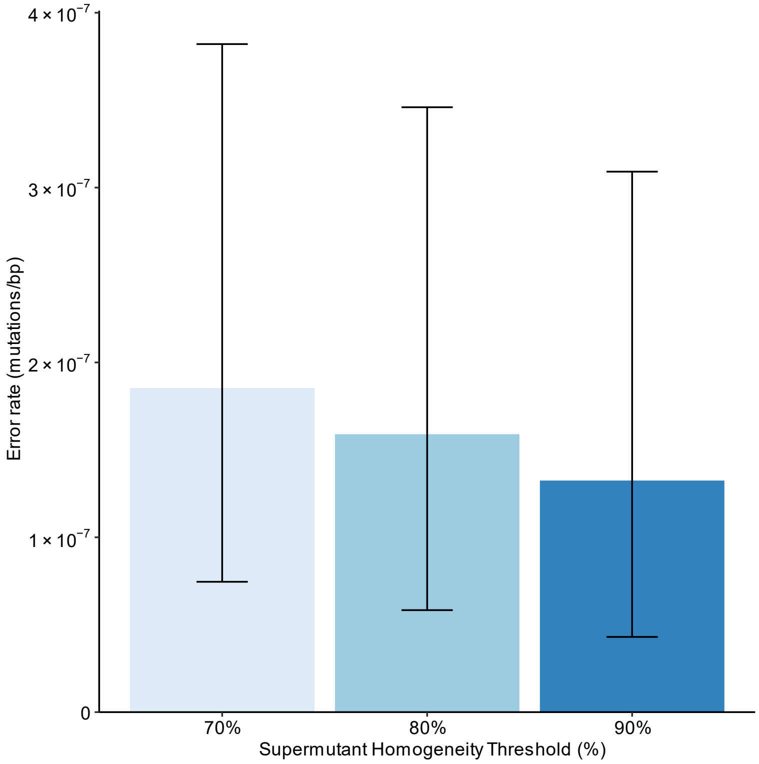
SaferSeqS libraries were constructed from a total of 74 plasma samples from cancer patients and subsequently evaluated with a multiplex PCR assay targeting recurrently mutated positions in human cancers and with tumor-specific single amplicon PCR assays.  The concordance between the two target enrichment methods is shown in the confusion matrix above.  In accordance with stochastic sampling effects, the two discordant plasma samples each harbored one supercalimutant when evaluated with either assay (**Supplementary Table 6**).



**Supplementary Figure 7**

**Performance of 182 primer pairs.**

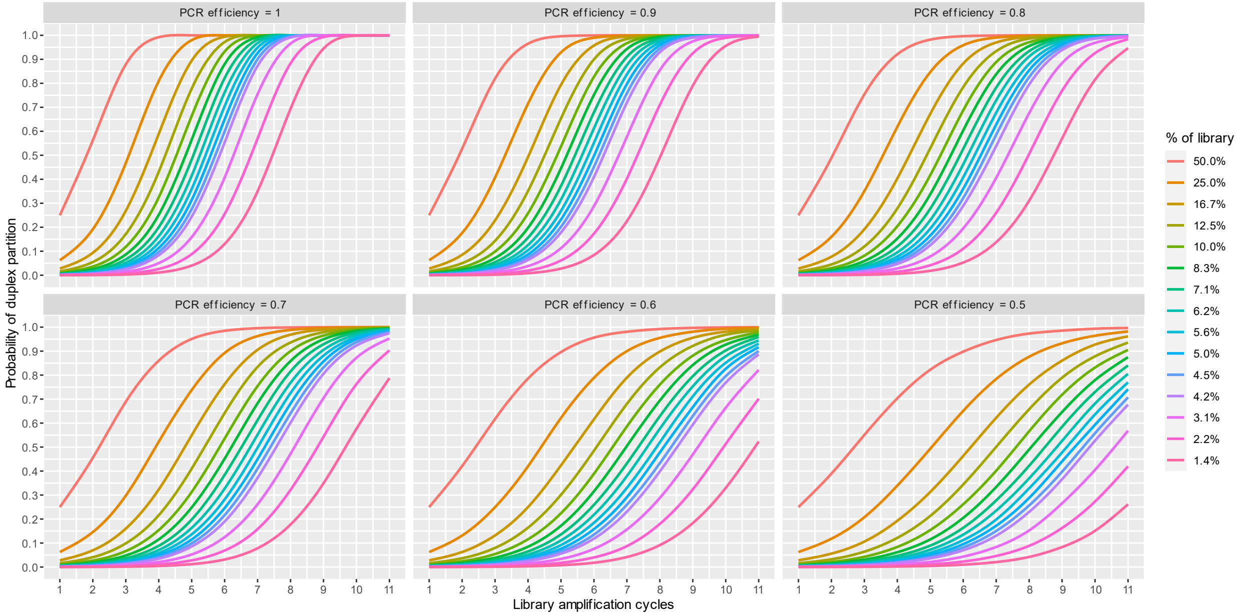
The proportion of on-target reads (i.e. the fraction of total reads that map to the intended target) for each of 182 SaferSeqS primer pairs tested to date. Of these 182 pairs, 163 (90%) exhibit an on-target rate of greater than 50%. The results presented reflect a single attempt at primer design. (Top) Distribution of the on-target rates for the 182 primer pairs. (Bottom) On-target rates for each of the individual 182 amplicons grouped by gene.



**Supplementary Figure 8**

**Effect of the supermutant threshold on SaferSeqS error rate.**

Reanalysis of the sequencing data from the mixing experiment presented in **Fig. 2** using supermutant thresholds of 70%, 80%, and 90%. The background mutation rates in all three cases were statistically indistinguishable from one another (P = 0.85, two-sided Z test for proportions), demonstrating that the performance of SaferSeqS is relatively invariant to the supermutant threshold used for scoring a UID family as a supercalimutant. Bars represent point estimates of background mutation rates using supermutant thresholds of 70%, 80%, and 90%. A total of 37,747,721, 37,747,670, and 37,747,476 bases were queried, and a total of 7, 6, and 5 supercalimutants were observed using supermutant thresholds of 70%, 80%, and 90%, respectively. Error bars represent exact 95% binomial confidence intervals of these point estimates.



## Supplementary Figure 9

### Effects of PCR efficiency and cycle number on duplex recovery.

The probability of recovering both strands of the original DNA duplexes (y-axis) is plotted against library amplification cycle number (x-axis).  Each pane in the figure represents the assumed PCR efficiency denoted at the top of the pane.  The proportion of the library amplification product used in the strand-specific PCRs are shown as colored curves, as specificied at the right of the figure.  Library amplification cycle number was varied from one to 11.  PCR efficiency was varied from 100% to 50% in 10% increments.  The proportion of library amplification product using in each strand-specific PCRs was varied from 50% to 1.4%. Probabilistic modeling was performed as described in the **Supplementary Note**.

## Supplementary Note

### Library construction

To address inefficiencies associated with library construction, we designed a strategy that relies on the sequential ligation of adapter sequences to the 3' and 5' DNA fragment ends[1] and the generation of double stranded molecular barcodes *in situ* (**Fig. 1a**). After DNA ends were dephosphorylated and repaired (**Fig. 1a**, step 1), an adapter was attached to the 3' end of DNA fragments (**Fig. 1a**, step 2). The adapter was a partially double-stranded DNA fragment with end modifications that selectively ligated to the 3' DNA ends and prevented adapter-dimer formation. Specifically, this adapter consisted of one oligonucleotide containing a 5' phosphate end modification (**Supplementary Table 7**, 3' N14 Adapter Oligo #1) which was hybridized to another oligonucleotide containing a 3' blocking group and deoxyuridines substituted for deoxythymidines (**Supplementary Table 7**, 3' N14 Adapter Oligo #2). This design permitted the use of adapters at high concentration in the ligation reaction which promoted efficient attachment to the 3' ends without the risk of significant dimer or concatemer formation[1]. Furthermore, the adapter contained a stretch of 14 random nucleotides in one of the two oligonucleotides which compromised one strand of the duplex UID. This step-wise sequence of reactions creates a cohesive end for efficient ligation of a second, 5' adapter. Following ligation of the 3' adapter, a second adapter (**Supplementary Table 7**, 5' Adapter) was ligated to the 5' DNA fragment ends via a nick translation-like reaction consisting of a DNA polymerase, cohesive end-specific ligase, and uracil DNA glycosylase (**Fig. 1a**, step 3). The concerted action of these enzymes synthesized the complementary strand of the UID, degraded the blocking portion of the 3' adapter, and ligated the extended adapter to the 5' DNA fragment end. The *in situ* generation of double stranded molecular barcodes uniquely barcoded each DNA fragment and obviated the need to enzymatically prepare duplex adapters, which has been noted to adversely affect input DNA recovery[2] (likely because the ends of enzymatically-prepared duplex adapters are less suitable substrates for ligation than the ends of adapters that are prepared chemically). Finally, the adapter-ligated fragments were subjected to a limited of number of PCR cycles to create redundant copies (UID "families") of the two original DNA strands (**Fig. 1a**, step 4).

### Effects of library amplification cycle number and efficiency

The number of PCR cycles and the efficiency of duplication during library amplification are critical SaferSeqS parameters. Because SaferSeqS relies on the partitioning of redundant Watson and Crick strand-derived copies into specific strand-specific PCRs for target enrichment, a requisite number of copies must be generated to ensure a high probability of duplex recovery. For example, assuming 100% efficiency, after one PCR cycle, each template DNA duplex is converted into two double stranded copies (one representing each strand), and there is only a 25% probability of properly distributing these two copies such that the one Watson strand-derived copy is partitioned into the Watson-specific PCR and the one Crick strand-derived copy is partitioned into the

Crick-specific PCR. Increasing the number of PCR cycles, or increasing the amplification efficiency, generates more redundant copies which in turn increases the probability of recovering the original DNA duplex.

We developed a probabilistic model to estimate the number of PCR cycles and amplification efficiency necessary for efficient duplex recovery. This model consisted of three steps: 1) simulate the number of PCR progeny generated during library amplification; 2) randomly partition these PCR copies into Watson and Crick strand-specific reactions; and 3) determine the duplex recovery—that is, the proportion of original DNA duplexes which have at least one Watson strand-derived copy partitioned into the Watson strand-specific reaction(s) and at least one Crick strand-derived copy partitioned into the Crick strand-specific reaction(s).

The number of PCR copies of the original template strands generated during each library amplification cycle follows a binomial distribution[3]. For the first PCR cycle, the number of strand-specific copies were initialized to one. It should be noted that the counts were initialized to one (instead of two) because the first library amplification cycle merely serves to denature the two original template strands and convert them into physically distinct double stranded forms. During the subsequent $i$th cycles of PCR, each of the $n_i$ PCR copies can replicate with probability $p$ (i.e. the efficiency of amplification) to generate a total of $n_{i+1}$ PCR copies equal to $n_i$ + Binom($n_i$, $p$). This process was iteratively repeated to simulate the number of progeny generated after $i$ PCR cycles. Formally, the number of total PCR copies generated can be expressed as follows:

$$n_i = \sum_{j=1}^{i-1} \mathrm{Binom}(n_j, p)\,;\ n_1 = 1$$

After library amplification, each original DNA duplex has been amplified to generate $n_{i,\mathrm{W}}$ copies of the Watson strand and $n_{i,\mathrm{C}}$ copies of the Crick strand as described above. Each of the $n_{i,\mathrm{W}}$ and $n_{i,\mathrm{C}}$ copies are randomly partitioned into Watson and Crick strand-specific PCR reactions with a probability $q$ that is equal to the fraction of the library used for each reaction. When the library is divided into a single Watson and single Crick strand-specific PCR, $q$ equals 50%. If the library is divided into two Watson and Crick strand-specific PCRs, $q$ equals 25%. The number of PCR copies that are partitioned into the appropriate strand-specific PCR ($N_{k,\mathrm{W}}$ or $N_{k,\mathrm{C}}$ for the $k$th Watson-specific or Crick-specific PCR, respectively) is drawn from a Binomial distribution with $n_{i,\mathrm{W}}$ or $n_{i,\mathrm{C}}$ "trials" and probability $q$ of "success" for the Watson and Crick copies, respectively. Therefore, the probability of partitioning at least one Watson-derived PCR copy into the $k$th Watson-specific PCR reaction is:

$$P\left(N_{k,W} > 0\right) = 1 - (1 - q)^{n_{i,W}}$$

Similarly, the probability of partitioning at least one Crick-derived PCR copy into the $k$th Crick-specific PCR reaction is:

$$P\left(N_{k,C} > 0\right) = 1 - (1 - q)^{n_{i,c}}$$

Both strands of an original DNA duplex can only be recovered if $N_{k,\mathrm{W}}$ and $N_{k,\mathrm{C}}$ are greater than zero. Because the partitioning of the PCR progeny is independent, the probability of duplex recovery is therefore predicted to be:

$$P\left(N_{k,W} > 0, N_{k,C} > 0\right) = [1 - (1 - q)^{n_{i,W}}][1 - (1 - q)^{n_{i,C}}]$$

We varied the PCR efficiency from 100% to 50%, the number of library amplification cycles from 1 to 11, and the fraction of the library used for each reaction from 50% to 1.4%. For each condition, we conducted 10,000 simulations of the above described process and report the average duplex recovery in **Supplementary Fig. 9**.

Multiplexing

SaferSeqS permits two types of multiplexing, one in which multiple targets are assayed in separate PCR reactions, and another in which multiple targets are assayed in the same PCR reaction. Because redundant Watson and Crick strand-derived copies are created during library amplification, the library should theoretically be able to be partitioned into multiple PCR reactions without adversely impacting recovery of the initial template molecules. For example, assuming a PCR efficiency of 70%, up to 22 targets can, in theory, be separately assayed with < 10% loss in recovery if a DNA library is amplified with 11 PCR cycles (**Supplementary Fig. 9**). In practice, we assayed either 100% or 4.4% of a library. The on-target rate was similar whether using 100% of 4.4% of the library, with 82% and 92% of reads properly mapping to the intended region. The number of duplex families recovered was also similar, with 7,825 and 6,769 recovered in the 100% and 4.4% library partitions, respectively.

Fragment size and recovery with anchored hemi-nested PCR

Anchored hemi-nested PCR[4] theoretically demonstrates a higher recovery of template molecules than traditional amplicon PCR. In traditional amplicon PCR, a template molecule must contain the both forward and reverse primer binding sites and the intervening sequence that defines the amplicon. In contrast, in anchored hemi-nested PCR, the template molecules must only harbor the union of the two gene-specific primer binding sites in order to be recovered. The combined footprints of the nested gene-specific primers used in SaferSeqS are approximately 30 bp, whereas the amplicon lengths employed by SafeSeqS for profiling cfDNA are typically 70-80 bp. Formally, assuming uniformly random fragment start/end coordinates, the probability of recovering a template molecule of length $L$ is $\frac{L-r}{L}$ where $r$ is the amplicon length in the case of traditional PCR or the length of the combined footprint of the gene specific primers in the case of anchored hemi-nested PCR. Thus, for cell-free DNA fragments of size ~167 bp[5], anchored hemi-nested PCR can theoretically recover ~25% more of the original template fragments than traditional amplicon PCR. Furthermore, unlike traditional amplicon PCR which produces predefined product sizes of that are dictated by the positions of the forward and reverse primers, anchored hemi-nested produces fragments of varying lengths with only one of the fragment ends dictated by the positions of the gene specific primers. Assuming template molecules of length *L* with uniformly random start/end coordinates, the observed fragment length after anchored hemi-nested PCR will be $\frac{L-r}{2}$ where *r* is the length of the combined footprint of the gene specific primers.

SaferSeqS bioinformatic pipeline

Reads were processed and mapped as described in **Methods**. The Watson and Crick reads for each sample were merged into a single BAM file and sorted by read name using SAMtools[6] so that mate pairs could be readily extracted. Custom Python scripts were used for subsequent reconstruction of the duplex families and identification of Watson supermutants, Crick supermutants, and supercalimutants.

First, reads were grouped into UID families while taking note of which reads were derived from the Watson and Crick strand by examining the value of their bitwise flag (i.e. FLAG field). Reads containing bitwise flagwise values of 99 and 147 are derived from the Watson strand and those containing bitwise flags of 83 and 163 are derived from the Crick strand. Reads with any other bitwise flag values were excluded from subsequent analysis.

Second, two additional quality control criteria were imposed during UID family grouping to ensure accurate determination of the endogenous molecular barcode (i.e. fragment end coordinate): 1) reads with soft clipping at the 5' or 3' of the fragment ends were excluded, 2) reads were required to contain the expected constant tag sequence (GCCGTCGTTTTAT) immediately following the exogenous UID with no more than one mismatch.

Third, because the number of possible exogenous UID sequences greatly exceeds the number of starting template molecules, "barcode collisions" in which two molecules share the same exogenous UID sequence but have different endogenous UIDs should be exceedingly rare. Specifically, the expected number of barcode collisions can be calculated from the classical "birthday problem" and is:

$$E[X] = n\left\{1 - \left(1 - \frac{1}{N}\right)^{n-1}\right\}$$

where *n* is equal to the number of template molecules and *N* is equal to the number of possible barcodes. For a 14 bp exogenous UID sequence (comprising a total of 268,435,456 possible sequences) and 10,000 genome equivalents, the expected number of collisions is 0.37, or 0.0037% of the input. We therefore required that each exogenous UID sequence could only be associated with one endogenous UID. In instances where an exogenous UID was associated with more than one endogenous UID, the largest family was preserved and all others were discarded.

Finally, because the exogenous barcodes themselves are susceptible to PCR and sequencing errors, we error-corrected UID sequences and regrouped the UID families using the UMI-tools network adjacency method[7].

After the reads were assembled into UID families, Watson supermutants, Crick supermutants, and supercalimutants were called as described in **Methods**. To exclude common polymorphisms, we excluded known germline mutations and all mutations in the Genome Aggregation Database (gnomeAD)[8] present at a population allele frequency greater than 0.01%. Reads comprising supercalimutants were subjected to a final manual inspection to exclude possible alignment artifacts.

Lower limit of detection of mutations with SaferSeqS

When applied to DNA from leukocytes or cell-free DNA of normal individuals, supercalimutants are found approximately once every five to ten million bp sequenced, which represents a lower limit to the sensitivity of the method for DNA from blood. It does not represent the method's analytical limit of detection, which could be considerably lower. The reason is that the supercalimutants found in these experiments could be due to mutations legitimately present in the starting templates rather than due to errors introduced by the SaferSeqS method or the sequencing. Blood cells continually divide during life, and the frequency of mutations we observe is consistent with estimates of non-clonal somatic mutation rates in healthy cells[9-11]. Moreover, other duplex sequencing methods[12,13] have reported mutation frequencies in normal tissues similar to those observed with SaferSeqS. Thus, the 100-fold reduction in error rate reported here may be an underestimate of the true error-correction capability of SaferSeqS, and the specificity of SaferSeqS to detect mutations may be limited by biological processes rather than by technical noise.

Whole genome sequencing studies of *in vitro* clonally expanded normal hematopoietic stem cells[9-11] have demonstrated that mutations in human blood progenitor cells accumulate at a rate of 14.2 per genome per year. The DNA used in this study for the mixture experiments was obtained from a set of individuals of average age 30. As a result, the expected frequency of non-clonal somatic single base substitutions in these samples is 426 per diploid genome, or approximately 7 × $10^{-8}$ mutations per bp. In this study we evaluated a total of 41,321,151 bases with SaferSeqS from DNA derived from healthy control subjects. Among these 41,321,151 bases, we detected 5 single base substitution supercalimutants, representing a mutation frequency of 12 × $10^{-8}$. To determine whether the frequency of supercalimutants observed is in accordance with previous estimates of non-clonal somatic mutation rates in healthy blood cells, we calculate the following exact one-sided binomial p-value:

$$P(X \geq 5) = 1 - \sum_{k=0}^{4} \binom{41{,}321{,}151}{k} (7 * 10^{-8})^{k} (1 - 7 * 10^{-8})^{41{,}321{,}151-\mathrm{k}} = 0.17$$

Therefore, there is no statistically significance difference between the number of supercalimutants observed and the predicted number of age-associated non-clonal somatic mutations arising from healthy hematopoietic stem cells.

Supplementary Note References

1 Makarov, V. & Laliberte, J. Enhanced Adapter Ligation. United States patent 10,208,338 B2 (2019).

2 Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* **34**, 547-555, doi:10.1038/nbt.3520 (2016).

3 Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* **43**, e143, doi:10.1093/nar/gkv717 (2015).

4 Zheng, Z. *et al.* Anchored multiplex PCR for targeted next-generation sequencing. *Nat Med* **20**, 1479-1484, doi:10.1038/nm.3729 (2014).

5 Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57-68, doi:10.1016/j.cell.2015.11.050 (2016).

6 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

7 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).

8 Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210, doi:10.1101/531210 (2019).

9 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

10 Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473-478, doi:10.1038/s41586-018-0497-0 (2018).

11 Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep* **25**, 2308-2316 e2304, doi:10.1016/j.celrep.2018.11.014 (2018).

12 Schmitt, M. W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12**, 423-425, doi:10.1038/nmeth.3351 (2015).

13 Salk, J. J. *et al.* Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Rep* **28**, 132-144 e133, doi:10.1016/j.celrep.2019.05.109 (2019).