GENETICS

Single-molecule, quantitative detection of low-abundance somatic mutations by high-throughput sequencing

Alexander Y. Maslov[1,2]*, Sergey Makhortov[3], Shixiang Sun[1], Johanna Heid[1], Xiao Dong[1,4], Moonsook Lee[1], Jan Vijg[1,5]*

Postzygotic somatic mutations have been found associated with human disease, including diseases other than cancer. Most information on somatic mutations has come from studying clonally amplified mutant cells, based on a growth advantage or genetic drift. However, almost all somatic mutations are unique for each cell, and the quantitative analysis of these low-abundance mutations in normal tissues remains a major challenge in biology. Here, we introduce single-molecule mutation sequencing (SMM-seq), a novel approach for quantitative identification of point mutations in normal cells and tissues.

INTRODUCTION

Mutations in the genome of somatic cells of multicellular organisms are the inevitable consequence of errors during DNA repair or replication. Somatic mutations cause cancer and have been implicated in other pathologies (*1*). Attempts have been made in the past to develop assays for the quantitative analysis of various types of mutations in cells and tissues (*2*–*5*). In view of the marked progress of DNA sequencing, one would think that somatic mutations should be easy to detect quantitatively in human or animal cells and tissues. In a very short time, an enormous amount of information has become available about somatic mutations in human tumors. However, tumors are clonal lineages with many mutations shared between the individual cells of the tumor. Mutations in normal tissues, however, are mostly unique for each cell, and their detection by sequencing remains a challenge because somatic mutations occur at low abundance and are spread through the reads, indistinguishable from sequencing errors. One way to overcome this problem is using a single cell–based approach (*6*, *7*). However, while the single-cell approach is currently the only method allowing comprehensive genome-wide assessment of somatic mutational loads, this method is resource- and time-consuming with a high price tag, which limits its broad application. An alternative approach, duplex sequencing (Duplex-Seq), is based on a comparative analysis of the complementary DNA strands and allows accurate quantitative identification of ultrarare somatic single-nucleotide variants (SNVs) in bulk DNA (*8*). While less demanding technically than single-cell sequencing, Duplex-Seq's capacity to suppress errors is limited to the square of the probability of errors on one strand. Here, we introduce single-molecule mutation sequencing (SMM-seq) for the accurate cost-effective assessment of somatic SNVs in bulk DNA extracted from normal cells and tissues.

RESULTS

SMM-seq library preparation

The key feature of SMM-seq is a two-step library preparation protocol (Fig. 1). First, rolling circle–based linear amplification (RCA) is used to produce single-stranded DNA molecules composed of multiple concatemerized copies of equally represented DNA strands of each particular DNA fragment. The amplification is carried out using an artificial thermostable polymerase having a strong strand displacement activity (*9*). This allows multiple cycles of denaturation-annealing-extension to ensure efficient and less biased amplification in a reaction we termed pulse-RCA. Because all these copies are independent replicas of the original DNA fragment, potential errors of amplification remain unique for each copy and do not propagate further. Copies of opposite strands are in an end-to-end orientation and separated by common spacers used as polymerase chain reaction (PCR) priming sites during the second step of the process when concatemerized copies are individually amplified and converted into a sequencing library (Fig. 1A). Thus, the resulting sequencing library is composed of PCR duplicates of multiple independent copies of an original DNA fragment assembled in rolling circle (RC) amplicons.

SMM-seq data analysis and variant calling

Sequencing reads originating from the same fragment are recognized on the basis of unique molecular identifiers (UMIs) introduced as part of hairpin-like adapters during library preparation. UMI families composed of reads originating from both strands of the original fragments are then used to identify the consensus sequence of each fragment. Consensus calls different from the corresponding positions on the reference genome are compared with a list of single-nucleotide polymorphisms (SNPs) of this particular DNA sample as well as with SNP database (dbSNP). This allows to filter out germline variants and identify potential de novo somatic mutations. A list of germline SNPs is obtained by analysis of conventional sequencing data of the same DNA sample performed in parallel with SMM-seq. The resulting list of potential somatic SNVs is further filtered to exclude low-confidence candidates and then saved for further analysis (Fig. 1B).

To determine the optimal analysis parameters, we assessed the frequency of somatic SNVs detected by the SMM variant–calling pipeline as a function of strand family size, i.e., the number of reads

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA. [2] Laboratory of Applied Genomic Technologies, Voronezh State University of Engineering Technologies, Voronezh, Russia. [3] Department of Programming and Information Technology, Voronezh State University, Voronezh, Russia. [4] Institute on the Biology of Aging and Metabolism and Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA. [5] Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China.

*Corresponding author. Email: alex.maslov@einsteinmed.edu (A.Y.M.); jan.vijg@einsteinmed.edu (J.V.)

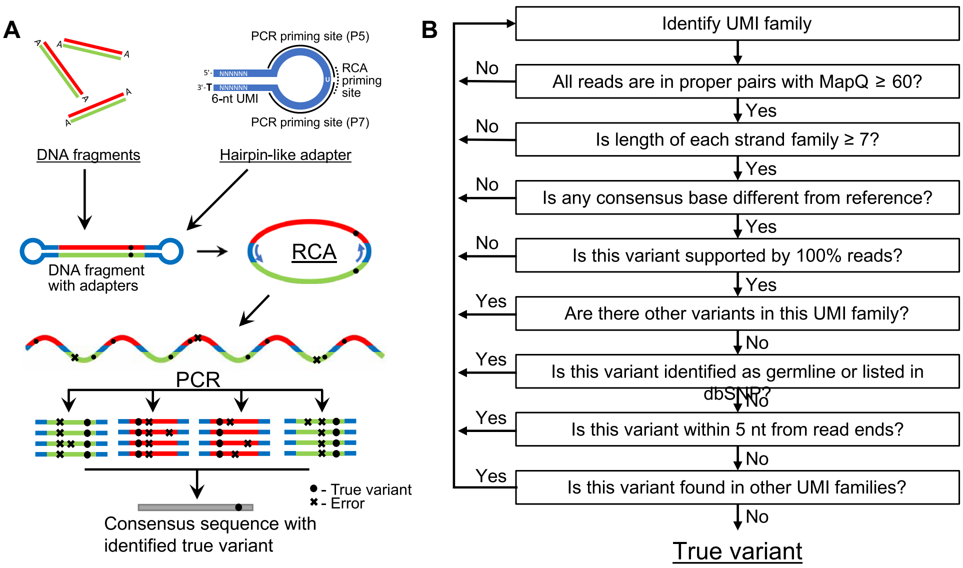
A
PCR priming site (P5)
5'- NNNNNN
3'-T NNNNNN
6-nt UMI
RCA priming site
PCR priming site (P7)
DNA fragments
Hairpin-like adapter
DNA fragment with adapters
RCA
PCR
● - True variant
✖ - Error
Consensus sequence with identified true variant
B
Identify UMI family
All reads are in proper pairs with MapQ ≥ 60?
Is length of each strand family ≥ 7?
Is any consensus base different from reference?
Is this variant supported by 100% reads?
Are there other variants in this UMI family?
Is this variant identified as germline or listed in dbSNP?
Is this variant within 5 nt from read ends?
Is this variant found in other UMI families?
Yes
No
True variant

Fig. 1. Outline of SMM-seq workflow and variant calling algorithm. (**A**) Both ends of end-repaired and A-tailed DNA fragments are ligated with a hairpin-like adapter. The adapter contains a 6–nucleotide (nt) long unique molecular identifier (UMI) in its stem part allowing identification of sequencing reads from the same original DNA fragment (UMI family) as well as identification of strand families. The hairpin-like adapter contains uracil in its loop part, allowing uracil-DNA glycosylase–mediated breakage and polymerase chain reaction (PCR) amplification when a conventional sequencing library is needed. The resulting dumbbell-like constructs, with intact uracils, serve as templates for the subsequent pulse-RCA reaction. Because each of the two ligated adapters contains an RCA priming site, the reaction starts from both sides of a fragment, generating identical products and ensuring higher efficiency of linear amplification. Single-stranded DNA contigs are then PCR-amplified to obtain multiple independent replicates of the original DNA fragments. Sequencing reads are aligned to the corresponding reference genome, UMI families identified and somatic variants are identified according to the computational algorithm shown (**B**). dbSNP, single-nucleotide polymorphism database.

representing each strand in a UMI family. We reasoned that each variant detected by SMM-seq is falling into one of the following categories—true positive (TP) or false positive (FP). Then, mutation frequency is a sum of frequencies of TP and FP, i.e., (TP + FP)/number of analyzed bases. The SMM library contains PCR replicates of multiple independent RC copies of each strand of the original fragments. The size of the UMI strand families (shown in green and red in Fig. 1) determines the chance of multiple PCR duplicates of the same RCA error. Thus, greater UMI strand families are less prone to FP calls. Hence, the frequency of FP mutations should decline with increasing family size. To test this, we determined the mutation frequency at different family size and normalized the results to the mutation frequency at a strand family size of two, as has been used in nanorate sequencing (NanoSeq), a modified version of Duplex-Seq (*10*). We found that increasing the minimum required strand family size from two to seven led to a statistically significant decrease in observed mutation frequency at each iteration, resulting in a more than twofold decrease (54% change). The excess mutations were considered artifacts and rejected. Further increasing strand family size no longer led to a statistically significant declines in detected mutation frequency (less than 10% change) (Fig. 2A). Thus, we used a cutoff level of seven reads per strand family as a qualifying criterion for variant calling at high accuracy.

Detection of induced SNVs by SMM-seq

As a proof of principle, we first performed SMM-seq analysis of DNA extracted from normal human IMR90 fibroblasts subjected in vitro to a single treatment with two different doses of *N*-ethyl-*N*-nitrosourea (ENU), a potent point mutagen (*6*). Here, we used sublethal doses of ENU that do not cause any noticeable cell death on IMR90 cells. Analysis of SMM-seq data revealed ~200 million positions per sample genome on average suitable for variant calling, i.e., the equivalence of ~7% of the genome. A regular sequencing library from IMR90 DNA was prepared, sequenced, and analyzed in parallel with SMM-seq to obtain a list of IMR90-specific germline SNPs. We found that our SMM-seq assay allows detection of mutagenic effects of ENU in all tested conditions (Fig. 2B and table S2). The lowest dose of ENU (25 μg/ml) increased the mutation frequency in IMR90 cells from 0.21 ± 0.02 to 0.36 ± 0.04 SNV/1 Mbp ($P = 0.005$), while ENU at 50 μg/ml led to a more than twofold increase of mutation frequency (0.54 ± 0.03 SNV/1 Mbp; $P = 9.7 \times 10^{-5}$). We also tested mutation spectra of somatic SNVs in control, nontreated cells and cells subjected to ENU treatment. We observed a distinct shift of mutational spectra upon ENU treatment, with the relative representations of TA/AT and TA/CG mutations [specific for ENU (*11*)] >2 times larger than in the untreated control cells (Fig. 2C and fig. S1). Thus, SMM-seq is capable of detecting somatic SNVs induced by low doses of mutagen.

Detection of aging-associated SNVs by SMM-seq

Next, we tested whether SMM-seq is capable of detecting physiological mutation burdens in human tissues accumulated during aging. We took advantage of our recently published study on the age-related mutational load in human liver that was performed using the gold standard single cell–based approach (*12*) and reanalyzed the same samples using SMM-seq assay. We used the whole-genome sequences of bulk DNAs from each participant from that same study for subtracting germline SNPs. SMM-seq libraries were prepared from DNA samples extracted from liver tissue of three young (5 months, 16 months, and 18 years old) and three aged people (56, 61, and 77 years old) (table S1). Analysis of SMM-seq data in this experiment revealed ~770 million positions qualified for variant calling per

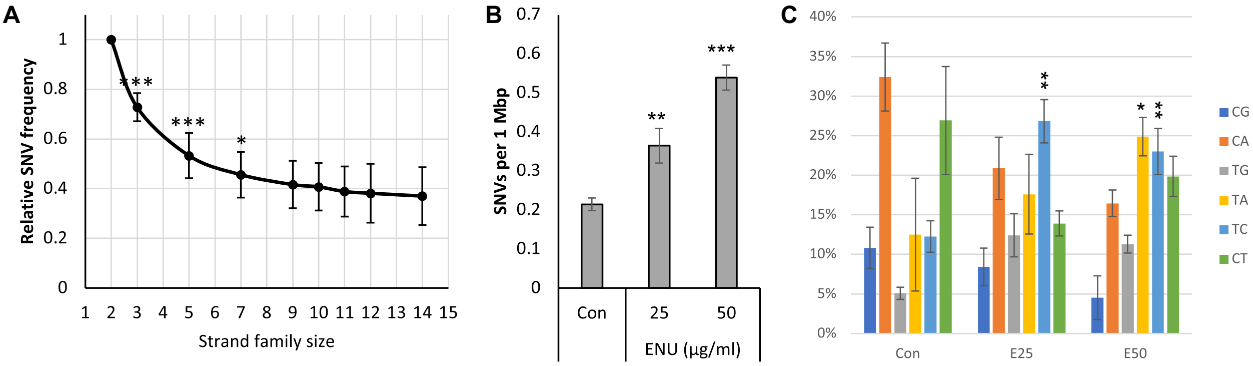

Fig. 2. Quantitative detection of induced somatic SNVs. (**A**) Relative mutation frequency as a function of strand family size. Statistical significance of difference with the previous value is shown. (**B**) Frequency of somatic SNVs in IMR90 cells 72 hours after treatment with different doses of ENU. (**C**) Relative representation of different mutation types in control cells and cells treated with ENU. Spectra of somatic SNVs in control cells and cells treated with ENU. All data points represent three biological replicates. Data are shown as average ± SD; asterisk (*) designates a statistically significant difference with its control (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

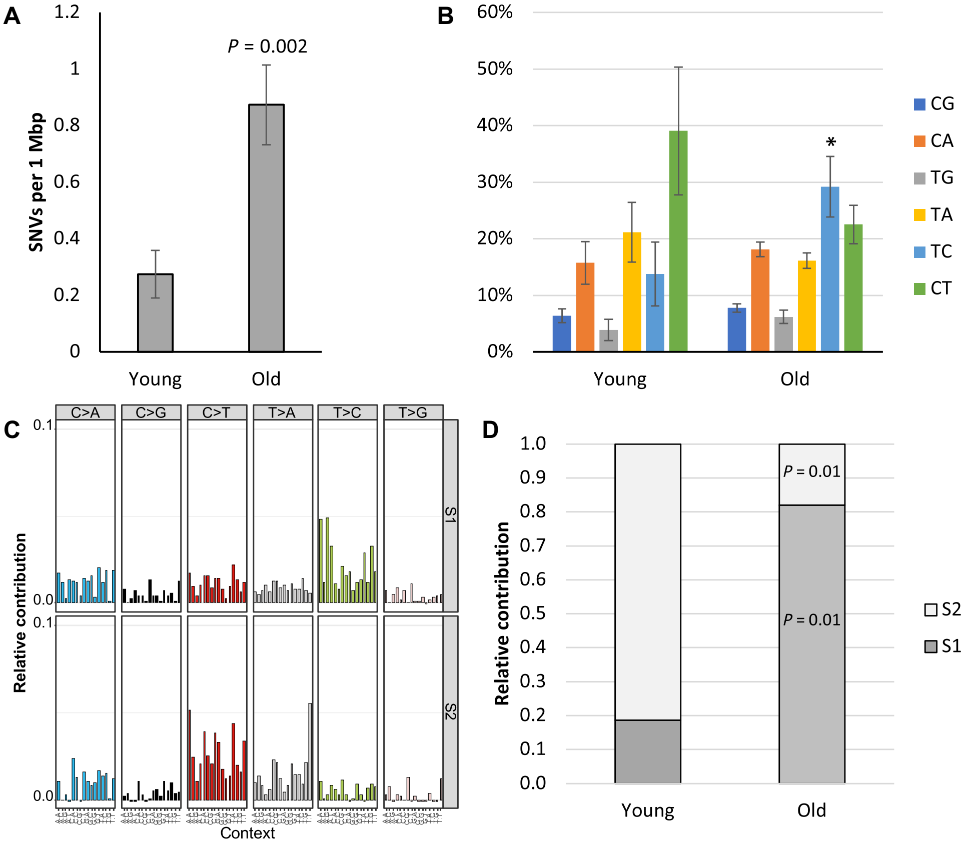

Fig. 3. Quantitative detection of somatic SNVs in normal human liver. (**A**) Frequency of somatic SNVs in liver of young and old participants. (**B**) Relative representation of different mutation types in liver of young and old participants. (**C**) Two mutational signatures de novo identified among variants detected by SMM-seq in the two different age groups. (**D**) Contributions of signatures S1 and S2 to somatic SNVs found in the liver of young and aged participants. All data points represent three biological replicates. Data are shown as average ± SD, asterisk (*) designates a statistically significant difference with its control ($P < 0.05$).

sample genome on average, the equivalent to ~26% of the sample genome. SMM-seq confirmed the age-related elevation in the somatic mutation frequency observed by the single-cell approach (Fig. 3A, fig. S2, and table S3). Mutation frequencies assayed by SMM-seq were 0.34 ± 0.09 and 0.96 ± 0.16 somatic SNVs/1 Mbp in the young and aged group, respectively ($P = 0.003$). Analysis of somatic SNV spectra revealed an almost twofold increase in the relative representation of TA to CG mutations in the liver DNA of aged people (16.2% in young versus 29.1% in aged) (Fig. 3B), similar to what has been observed by the single-cell approach.

Assessment of mutation signatures using SMM-seq data

To get further insight into the mutation spectra in the aged human liver, we performed nonnegative matrix factorization and extracted two de novo mutation signatures, S1 and S2 (Fig. 3C) (*13*, *14*), from the mutation spectra of six analyzed samples. Signature S1 was found to be substantially increased in the aged group ($P = 0.0134$; Fig. 3D and fig. S4) and associated with aging signature SBS5 (cosine similarity: 0.904, $P < 0.001$ by permutation test) (*15*). Signature S2 was dominant in the young group, with an abundance of CG to TA transitions and TA to AT transversions, but the source of signature S2 is not clear, and we did not find any substantial similarity between signature S2 and known COSMIC (Catalogue of Somatic Mutations in Cancer) signatures. These signature analysis results, as well as our results on the age-related increase of mutational load, are in good agreement with our previous findings (*12*). Thus, SMM-seq is capable of detecting somatic SNVs accumulated in normal human tissues under physiological conditions.

DISCUSSION

The various approaches using duplex consensus sequencing for the identification of rare mutations, i.e., the original Duplex-Seq (*8*), BotSeqS (*16*), and NanoSeq (*10*), are all based on analysis of the two opposite DNA strands to eliminate potential errors. The error rate of these approaches is determined by the probability of two complementary errors in both strands and can be defined as $P(E)^2$, where $P(E)$ is the probability of error on any of two strands. SMM-seq is not limited to two strands only since it uses sequencing data from multiple independent copies of each strand for variant calling. Conversely, SMM-seq's error rate can be calculated as $P(E)^N$, where N is the number of independent copies produced in the linear amplification step. Naturally, copies of the same strand cannot be distinguished from the sequencing data, but our results on variant calling using strand families of different sizes clearly demonstrated that the detected mutation frequency is plateauing at a family size of seven and further (Fig. 2A). This indicates that, at this size, each strand family contains descendants of more than one copy of the original DNA fragment and no further improvement of accuracy is possible. Notably, despite virtually unlimited accuracy in base calling on each strand, it is still necessary to have representatives of both to filter out possible artifacts produced by DNA damage, which are expected to be present on one strand only (Fig. 1B).

As we demonstrated, SMM-seq is capable of detecting both induced and naturally occurring somatic SNVs in normal human cells and tissues. The SMM-seq results are in line with results obtained using the single cell–based approach, currently the gold standard in the field. However, usage of SMM-seq is significantly less resource demanding. SMM-seq is more accurate than Duplex-Seq–based approaches because of the presence of multiple independent copies of the original DNA fragment. Thus, SMM-seq is a practical approach that, together with our previously developed Structural Variant Search (SVS) assay for detecting somatic structural variants (*17*, *18*), is well suited for the comprehensive assessment of genome integrity in large-scale human studies.

MATERIALS AND METHODS

Cell culture and treatment

Human normal lung IMR90 fibroblasts were maintained in 10% CO_2 and 3% O_2 atmosphere at 37°C in Dulbecco's modified Eagle's medium (GIBCO, Grand Island, NY, USA) supplemented with 10% fetal bovine serum (GIBCO). Twenty-four hours after cell seeding, the culturing media were changed for media containing different doses of ENU. Cells were harvested 72 hours after ENU was applied. Complete media supplemented with ENU (Sigma-Aldrich, St. Louis, MO, USA) were prepared immediately before application from stock solution (100 mg/ml in 100% ethyl alcohol). Control cells were cultured in the presence of the vehicle only.

Human specimens

Frozen human hepatocyte samples were purchased from Lonza Walkersville Inc. All six selected hepatocyte donors were healthy participants of various age and gender without any liver cancer or other liver pathology history (table S1).

DNA isolation

DNA from fibroblasts and hepatocytes was isolated using Quick-gDNA Blood MiniPrep (Zymo Research Corporation, Irvine, CA, USA) according to the manufacturer's instructions and quantified using a Qubit kit (Thermo Fisher Scientific, USA).

SMM library preparation and sequencing

Genomic DNA was first fragmented by double digestion with restriction endonucleases Alu I and Mlu CI (NEB, USA), overnight at 37°C. After purification using 1.5× AMPure XP beads (Beckman Coulter, USA), the fragmented DNA was further processed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). The adapter provided with the kit was replaced with custom adapter P5_HP6N. After double-sided size selection using AMPure XP beads (Beckman Coulter), resulting dumbbell-like product was quantified with the Qubit kit (Thermo Fisher Scientific) and analyzed on 2100 Bioanalyzer instrument with a High Sensitivity DNA kit (Agilent, USA). To prepare P5_HP6N adapter, 86 μl of 10 μM solution of oligonucleotide 5′-TCTTC TACAGT NNNNNN AGATCG GAAGAG CACACG TCTGAA CTCCAG TC /ideoxyU/ ACACTC TTTCCC TACACG ACGCTC TTCCGA TCT-3′ (IDT, USA) in 0.1× tris-EDTA (TE) buffer was first exposed to 95°C for 5′ followed by 37°C for 5′ to form a hairpin. Next, to fill in the extending 5′-end, the self-annealed oligonucleotide was supplemented with 10 μl of CutSmart buffer (NEB), 2 μl of 10 mM deoxynucleotide triphosphates (dNTPs) mix (NEB) and 10 U of Klenow Fragment (3′ → 5′ exo-) (NEB) and incubated at 37°C for 30′. After purification with the QIAquick Nucleotide Removal Kit (QIAGEN, USA), the hairpins were digested with 10 U of HpyCH4III (NEB) for 1 hour at 37°C, then purified again with the QIAquick Nucleotide Removal Kit (QIAGEN), and eluted with 100 μl of EB to obtain ready-to-use adapter solution.

Next, for SMM-seq library preparation, samples were diluted on the basis of assessed molar concentration. Assuming 150PE sequencing mode and 30 Gb of data per sample, the dilution coefficient (D) was calculated using the formula $D = \mathrm{M} \times \mathrm{N_A} \times 2/10^{25}$, where M is sample concentration (pM) and N_A is the Avogadro constant. Resulting suspension contains ~8 amol (5 million molecules) of DNA in 1 μl. Next, 1 μl of diluted sample was used as a template in pulse-RCA reaction. The pulse-RCA was performed in 20-μl reaction containing 1 μl of diluted sample, 1 μl of P5-RCA oligo (5′-GTAGGGAAAGAGTGTAGACTGGAGTTC-3′), 25 U (0.5 μl) of strand displacement polymerase HS (BIORON Diagnostics GmbH, Germany), 2 μl of buffer, 1 μl of 10 mM dNTPs mix (NEB), 0.6 μl of 100 mM $MgCl_2$, and 13.9 μl of water. The pulse-RCA program was set as follows: 92°C for 2 min (1); 92°C for 30 s (2); 60°C for 30 s (3); 65°C for 150 s (4); go to (3) nine times; hold at 4°C. Product of amplification reaction was purified with 1.5× AMPure XP beads and resuspended in 23 μl of TE buffer. The entire volume of RC amplification was PCR-amplified in 50-μl reaction volume containing 23 μl of RCA product, 25 μl of NEBNext Ultra II Q5 Master Mix, and 1 μl of P5 and P7 dual index oligos. The PCR program was set as follows: 98°C for 30 s (1); 98°C for 10 s (2); 65°C for 75 s (3); go to (2) eight times (4); 65°C for 5 min (5); 4°C forever. The PCR product was purified with 0.7X AMPure XP beads and resuspended in 30 μl of TE buffer. After quantification with Qubit, samples were pooled and sequenced on an Illumina NovaSeq instrument using 150 paired-end mode.

Conventional sequencing library was prepared by PCR amplification of adapter-ligated samples in 30-μl reaction volume containing 11 μl of undiluted ligated sample, 2 U of USER enzyme (NEB), 15 μl of NEBNext Ultra II Q5 Master Mix, and 1 μl of P5 and P7 dual index oligos. The PCR program was set as follows: 37°C for 15 min (1); 98°C for 30 s (2); 98°C for 10 s (3); 65°C for 75 s (4); go to (3) four times (4); 65°C for 5 min (5); 4°C forever. The PCR product was purified with 0.7× AMPure XP beads and resuspended in 30 μl of TE buffer. After quantification with Qubit, samples were pooled and sequenced on the Illumina NovaSeq instrument using 150 paired-end mode.

Data processing and variant calling

Raw sequence reads were adapter- and quality-trimmed, aligned to human reference genome, realigned, and recalibrated on the basis of known indels as we described previously (*7*) except that deduplication step was omitted.

For variant calling, we developed a set of filters that were applied to each position in SMM-seq data. Only reads in proper pairs, with mapping quality not less than 60 and without secondary alignments, were taken in consideration. Positions in SMM-seq data were considered as qualified for variant calling if it is covered by UMI family containing not less than seven reads from each strand and this position is covered at least 20× in regular sequencing data. The qualified position was considered as a potential variant if all the reads within a given UMI family reported the same base at this position and this base was different from the corresponding reference genome. Next, to filter out germline variants, we checked if a found potential variant is in a list of SNPs of this DNA sample as well as in dbSNP. A list of sample specific germline SNPs was prepared by analysis of conventional sequencing data with Genome Analysis Toolkit (GATK) HaplotypeCaller. Last, a variant was rejected if one or more reads of a different UMI family in SMM data or in conventional data contained the same variant. SNV frequency was calculated as a ratio of the number of identified variants to the total number of qualified positions.

Statistical analysis

Statistic tests were performed using Microsoft Office Excel (2013). All the experiments were performed in three biological replicates, and results are expressed as mean and SD. Statistical significance of differences between experimental groups was determined using two-tailed *t* test. The permutation test to calculate significance of cosine similarity was performed using an R package PharmacoGx (*19*) (cosinePerm; $n = 1000$).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abm3259

REFERENCES AND NOTES

1. J. Vijg, X. Dong, Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell* **182**, 12–23 (2020).
2. B. N. Ames, F. D. Lee, W. E. Durston, An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci. U.S.A.* **70**, 782–786 (1973).
3. J. McCann, E. Choi, E. Yamasaki, B. N. Ames, Detection of carcinogens as mutagens in the Salmonella/microsome test: Assay of 300 chemicals. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 5135–5139 (1975).
4. R. J. Albertini, K. L. Castle, W. R. Borcherding, T-cell cloning to detect the mutant 6-thioguanine-resistant lymphocytes present in human peripheral blood. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6617–6621 (1982).
5. J. Vijg, H. van Steeg, Transgenic assays for mutations and cancer: Current status and future perspectives. *Mutat. Res.* **400**, 337–354 (1998).
6. M. Gundry, W. Li, S. B. Maqbool, J. Vijg, Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic Acids Res.* **40**, 2032–2040 (2012).
7. X. Dong, L. Zhang, B. Milholland, M. Lee, A. Y. Maslov, T. Wang, J. Vijg, Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**, 491–493 (2017).
8. M. W. Schmitt, S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, L. A. Loeb, Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14508–14513 (2012).
9. K. B. Ignatov, E. V. Barsova, A. F. Fradkov, K. A. Blagodatskikh, T. V. Kramarova, V. M. Kramarov, A strong strand displacement activity of thermostable DNA polymerase markedly improves the results of DNA amplification. *Biotechniques* **57**, 81–87 (2014).
10. F. Abascal, L. M. R. Harvey, E. Mitchell, A. R. J. Lawson, S. V. Lensing, P. Ellis, A. J. C. Russell, R. E. Alcantara, A. Baez-Ortega, Y. Wang, E. J. Kwa, H. Lee-Six, A. Cagan, T. H. H. Coorens, M. S. Chapman, S. Olafsson, S. Leonard, D. Jones, H. E. Machado, M. Davies, N. F. Øbro, K. T. Mahubani, K. Allinson, M. Gerstung, K. Saeb-Parsy, D. G. Kent, E. Laurenti, M. R. Stratton, R. Rahbari, P. J. Campbell, R. J. Osborne, I. Martincorena, Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
11. C. W. Op het Veld, S. van Hees-Stuivenberg, A. A. van Zeeland, J. G. Jansen, Effect of nucleotide excision repair on hprt gene mutations in rodent cells exposed to DNA ethylating agents. *Mutagenesis* **12**, 417–424 (1997).
12. K. Brazhnik, S. Sun, O. Alani, M. Kinkhabwala, A. W. Wolkoff, A. Y. Maslov, X. Dong, J. Vijg, Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *Sci. Adv.* **6**, eaax2659 (2020).
13. F. Blokzijl, R. Janssen, R. van Boxtel, E. Cuppen, MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
14. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. M. Dermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
15. M. Petljak, L. B. Alexandrov, J. S. Brammeld, S. Price, D. C. Wedge, S. Grossmann, K. J. Dawson, Y. S. Ju, F. Iorio, J. M. C. Tubio, C. C. Koh, I. Georgakopoulos-Soares, B. Rodríguez-Martín, B. Otlu, S. O'Meara, A. P. Butler, A. Menzies, S. G. Bhosle, K. Raine, D. R. Jones, J. W. Teague, K. Beal, C. Latimer, L. O'Neill, J. Zamora, E. Anderson, N. Patel, M. Maddison, B. L. Ng, J. Graham, M. J. Garnett, U. M. Dermott, S. Nik-Zainal, P. J. Campbell, M. R. Stratton, Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
16. M. L. Hoang, I. Kinde, C. Tomasetti, K. W. McMahon, T. A. Rosenquist, A. P. Grollman, K. W. Kinzler, B. Vogelstein, N. Papadopoulos, Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9846–9851 (2016).
17. W. Quispe-Tintaya, T. Gorbacheva, M. Lee, S. Makhortov, V. N. Popov, J. Vijg, A. Y. Maslov, Quantitative detection of low-abundance somatic structural variants in normal cells by high-throughput sequencing. *Nat. Methods* **13**, 584–586 (2016).
18. W. Quispe-Tintaya, M. Lee, X. Dong, D. A. Weiser, J. Vijg, A. Y. Maslov, Bleomycin-induced genome structural variations in normal, non-tumor cells. *Sci. Rep.* **8**, 16523 (2018).
19. P. Smirnov, Z. Safikhani, N. el-Hachem, D. Wang, A. She, C. Olsen, M. Freeman, H. Selby, D. M. A. Gendoo, P. Grossmann, A. H. Beck, H. J. W. L. Aerts, M. Lupien, A. Goldenberg,

B. Haibe-Kains, PharmacoGx: An R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32**, 1244–1246 (2016).

Acknowledgments

Funding: This work was funded by the National Institutes of Health grant P01 AG017242 (to J.V.), National Institutes of Health grant U01 ES029519 (to J.V.), National Institutes of Health grant U01 HL145560 (to J.V.), National Institutes of Health grant P01 AG047200 (to J.V.), and Antonov's Foundation grant (to A.Y.M.). **Author contributions:** Conceptualization: A.Y.M. and J.V. Methodology: A.Y.M. and J.V. Investigation: A.Y.M., J.H., and M.L. Code writing: S.M. Data analysis and interpretation: A.Y.M., J.V., S.S., and X.D. Writing—original draft: A.Y.M. Writing—review and editing: J.V. and A.Y.M. **Competing interests:** A.Y.M. and J.V. are inventors on a patent pending related to this work filed by Albert Einstein College of Medicine (U.S. provisional application no. 63/277,955, filed 10 November 2021]. A.Y.M., X.D., M.L., and J.V. are cofounders of SingulOmics Corp. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Part of the sequencing data may be accessed at dbGaP - phs001956.v1.p1 (www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001956.v1.p1). Remaining data may be accessed using the following link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA758911. Analysis scripts may be accessed at https://zenodo.org/record/5804750#.Ycnt8S-B2-Y and https://github.com/msd-ru/SMM.

Submitted 26 September 2021
Accepted 18 February 2022
Published 8 April 2022
10.1126/sciadv.abm3259