RESEARCH ARTICLE SUMMARY

HUMAN GENETICS

RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues

Keren Yizhak, François Aguet, Jaegil Kim, Julian M. Hess, Kirsten Kübler, Jonna Grimsby, Ruslana Frazer, Hailei Zhang, Nicholas J. Haradhvala, Daniel Rosebrock, Dimitri Livitz, Xiao Li, Eila Arich-Landkof, Noam Shoresh, Chip Stewart, Ayellet V. Segrè, Philip A. Branton, Paz Polak, Kristin G. Ardlie, Gad Getz*

INTRODUCTION: Cancer genome studies have contributed to the analysis and discovery of somatic mutations that drive cancer growth. However, studying the genetic makeup of a tumor when it is already fully developed limits our ability to uncover how and which somatic mutations accumulate in normal tissues in the stages preceding cancer initiation. To address this challenge, recent studies performed deep sequencing in a limited number of tissue types and a small number of individuals, identifying a large number of microscopic clones carrying somatic mutations, some in known cancer genes. These findings emphasize the need to uncover the genomic events that occur in all normal tissues. Although efforts have begun to collect and analyze DNA from normal tissues, we still lack a comprehensive catalog of genetic events and clonal properties across a large number of tissues and individuals. By analyzing the information-rich content in RNA now available from recent advances in RNA sequencing methods, we may be able to substantially expand the scope and scale of these studies.

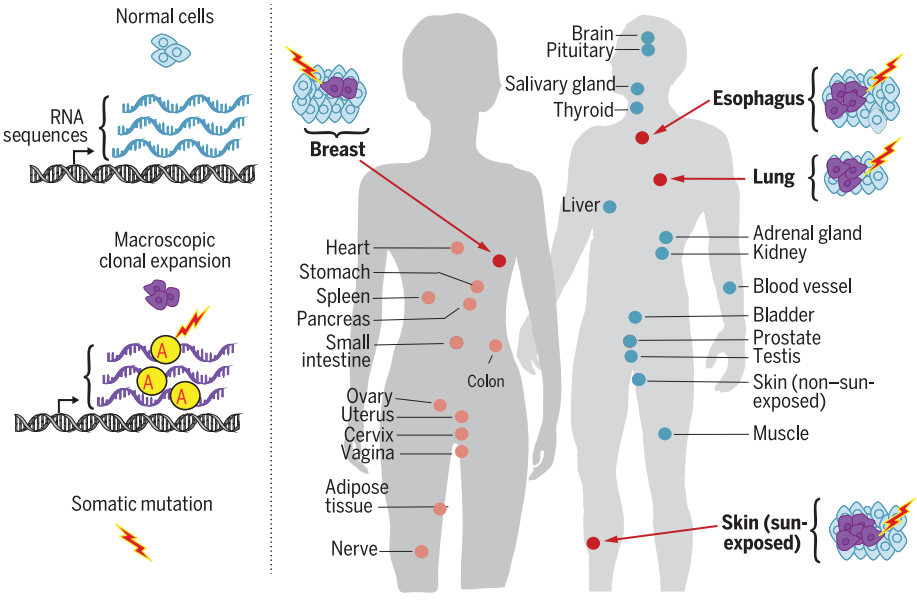
RATIONALE: Some mutations found in the DNA can be detected in the corresponding RNA, depending on the mutation allele fraction and sequence coverage. We therefore hypothesized that a careful analysis of RNA sequences from normal bulk tissues could uncover somatic mutations reflecting macroscopic clones within the samples. In this work, we used the large collection of RNA sequences from the Genotype–Tissue Expression (GTEx) project, representing more than 6700 samples from ~500 individuals, spanning across 29 different normal tissues.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aaw0726

RESULTS: We developed a new method, called RNA-MuTect, to identify somatic mutations using a tissue-derived RNA sample and its matched-normal DNA. We validated RNA-MuTect on both tumor-adjacent and cancer samples from The Cancer Genome Atlas (TCGA), wherein DNA and RNA were coextracted from the same samples. Focusing on mutations contained within sufficiently covered sequences, RNA-MuTect achieved high sensitivity and precision, enabling the discovery of most driver events and mutational processes from TCGA tumor RNA data. When applied to the GTEx dataset of normal tissues, multiple somatic mutations were detected in almost all individuals and tissues studied here, including in known cancer genes. The three tissues with the largest number of somatic mutations were sun-exposed skin, esophagus mucosa, and lung; this finding suggests that environmental exposure can promote somatic mosaicism. Both the individuals' age and tissue-specific proliferation rate were found to be associated with the number of detected mutations. A dN/dS (ratio of nonsynonymous to synonymous substitutions) analysis suggested that some of the mutations identified in cancer genes may confer a selective advantage. In addition, allelic imbalance events at the chromosome arm level were detected in normal tissues.

CONCLUSION: Genetic clones carrying somatic mutations are detected across normal tissues to different extents, and these differences depend on factors such as the tissue's exposure to environmental mutagens, natural architecture, proliferation rate, and the microenvironment. Some of these clones may be the result of genetic drift. Others, however, may develop as a result of positive selection driven by certain somatic events, thus potentially representing the earliest stages of tumorigenesis. Higher-resolution studies of normal tissues and precancerous lesions are required if we are to advance our understanding of both aging and early cancer development. ■



Somatic clonal expansions in normal human tissues. RNA sequences from 29 normal human tissues collected as part of the Genotype–Tissue Expression (GTEx) project are analyzed using RNA-MuTect, a method developed for detecting somatic mutations in RNA-seq data. Macroscopic clonal expansions, characterized by shared somatic mutations, are detected in all tissues; skin, esophagus, and lung have the largest number of somatic mutations.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: gadgetz@broadinstitute.org
Cite this article as K. Yizhak *et al.*, *Science* **364**, eaaw0726 (2019). DOI: 10.1126/science.aaw0726

RESEARCH ARTICLE

HUMAN GENETICS

RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues

Keren Yizhak[1], François Aguet[1], Jaegil Kim[1], Julian M. Hess[1], Kirsten Kübler[1,2,3], Jonna Grimsby[1], Ruslana Frazer[1], Hailei Zhang[1], Nicholas J. Haradhvala[1,2], Daniel Rosebrock[1], Dimitri Livitz[1], Xiao Li[1], Eila Arich-Landkof[1,2], Noam Shoresh[1], Chip Stewart[1], Ayellet V. Segrè[1,3,4], Philip A. Branton[5], Paz Polak[6], Kristin G. Ardlie[1], Gad Getz[1,2,3,7]*

How somatic mutations accumulate in normal cells is poorly understood. A comprehensive analysis of RNA sequencing data from ~6700 samples across 29 normal tissues revealed multiple somatic variants, demonstrating that macroscopic clones can be found in many normal tissues. We found that sun-exposed skin, esophagus, and lung have a higher mutation burden than other tested tissues, which suggests that environmental factors can promote somatic mosaicism. Mutation burden was associated with both age and tissue-specific cell proliferation rate, highlighting that mutations accumulate over both time and number of cell divisions. Finally, normal tissues were found to harbor mutations in known cancer genes and hotspots. This study provides a broad view of macroscopic clonal expansion in human tissues, thus serving as a foundation for associating clonal expansion with environmental factors, aging, and risk of disease.

As cells divide during life, they accumulate somatic mutations. Although most of these mutations are thought to be either neutral or slightly deleterious (*1*), a few may increase cellular fitness and contribute to clonal expansion. This process is associated with aging as well as with diseases such as coronary heart disease (*2*, *3*), neurological disorders (*4*), and cancer (*5*). In cancer, the accumulation of several mutations (known as "cancer drivers") eventually may transform the cells and promote uncontrolled cellular growth. Despite work contributing to our understanding of the molecular and cellular aspects of cancer (*6*–*14*), we still only partially understand the initiation and progression of this disease. Acknowledging this gap, studies have focused on studying somatic mutations in normal human tissues and precancerous lesions, aiming to identify early clonal expansions (*3*, *15*–*18*). Clonal expansions detected in normal blood are enriched with mutations in several genes implicated in hematologic cancers (*3*, *19*). Ultradeep sequencing studies by Martincorena *et al.* (*17*, *18*) in normal skin and esophagus tissues focused on 74 cancer genes and detected a high burden of low-allele frequency mutations associated with skin and esophagus squamous cell carcinoma. Despite these associations, it remains unclear which specific clones will eventually develop into cancer. Collectively, these findings emphasize the need to comprehensively map and study the prevalence and size of clonal expansion across human tissues.

A pipeline for detecting somatic mutations using RNA-seq data

For genomic data derived from normal tissues, we leveraged the Genotype–Tissue Expression (GTEx) project (*20*), a collection of data generated from more than 30 normal primary tissues from hundreds of healthy individuals. These data include RNA sequencing (RNA-seq) data of the tissues as well as whole-genome and whole-exome sequencing data of DNA extracted from matched blood samples (release V7), providing an opportunity to explore all genes and tissues for the existence of macroscopic clones that have expanded to a detectable level in bulk RNA-seq.

To detect somatic mutations from bulk RNA-seq data, we needed to first develop a pipeline, called RNA-MuTect, to analyze this type of data. To develop our approach for detecting somatic mutations from RNA-seq data, we initially focused on a training set of 243 tumor samples (representing six tumor types) from The Cancer Genome Atlas (TCGA) for which both DNA and RNA were co-isolated from the same cells (table S1). Applying our standard somatic mutation calling pipeline (that was developed for DNA) to both DNA and RNA from the tumor samples, and using the matched-normal DNA as a germline control (*21*), we found that the number of mutations in RNA exceeded the number in the corresponding DNA by a factor of 5 (Fig. 1A and fig. S1A) (*21*). Moreover, 65% of the DNA-based mutations were not detected in the RNA, and 92% of the RNA-based mutations were not found in the DNA (*21*). One obvious reason for not detecting DNA-based mutations in the RNA is the insufficient sequence coverage in genes with low expression levels; indeed, in a typical RNA sample, only 55% of the transcriptome had sufficient coverage (≥95% sensitivity) to detect mutations at the median DNA allele fraction (fig. S1B). When accounting for the actual allele fractions of the DNA mutations and coverage of RNA transcripts, RNA-MuTect detected 82% of the sufficiently covered mutations (fig. S2, A to D) (*21*).

Next, to address the excessive mutations detected only in the RNA, we developed RNA-MuTect, which is based on several key filtering steps (fig. S3), including (i) removal of alignment errors by using two different RNA aligners; (ii) removal of sequencing errors by a site-specific error model built upon thousands of normal RNA-seq data; and (iii) removal of RNA editing sites using known databases (*21*). The vast majority (93%) of RNA mutations were filtered out (fig. S2, E to G), reaching a median precision of 0.91 across samples (Fig. 1B), and only a median of three detected mutations per sample remained in the RNA-only set. RNA-MuTect retained a high overall median sensitivity of 0.7 after filtering (Fig. 1B and fig. S2E) (*21*), removing as few as 10% of mutations that were detected in the DNA. Of note, RNA-MuTect outperformed previous methods (*22*, *23*) in terms of both sensitivity and precision for detecting mutations in RNA-seq (*21*). To evaluate the robustness of RNA-MuTect on an independent dataset, we collected a validation set of 303 TCGA samples representing six tumor types (five differed from the training set; table S1). RNA-MuTect achieved high sensitivity and precision on the validation set, in agreement with the training set results (sensitivity of 0.72 and precision of 0.87; Fig. 1B).

The high overall performance of RNA-MuTect enabled us to apply our standard tools for finding drivers and mutational signatures to RNA-based mutations (*21*, *24*, *25*), which yielded results very similar to what was found in the DNA (Fig. 1, C and D, and figs. S4 and S5) (*21*). Our analysis did, however, identify a yet-unreported mutational signature in the RNA dominated by C>T mutations; this signature represented only 7% of the mutations, with the majority originating from a single colon cancer sample (Fig. 1D). Of these mutations, 75% were sufficiently covered but not detected in the DNA, which suggests that this signature may reflect a C>U RNA-editing process.

Notably, to obtain a conservative (i.e., higher) estimate of the false positive rate, we considered

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Ocular Genomics Institute, Department of Ophthalmology, Massachusetts Eye and Ear, Boston, MA, USA. [5]Biorepositories and Biospecimen Research Branch, Cancer Diagnosis Program, National Cancer Institute, Bethesda, MD, USA. [6]Oncological Sciences, Icahn School of Medicine at Mount Sinai Hospital, New York, NY, USA. [7]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA.
***Corresponding author. Email: gadgetz@broadinstitute.org**

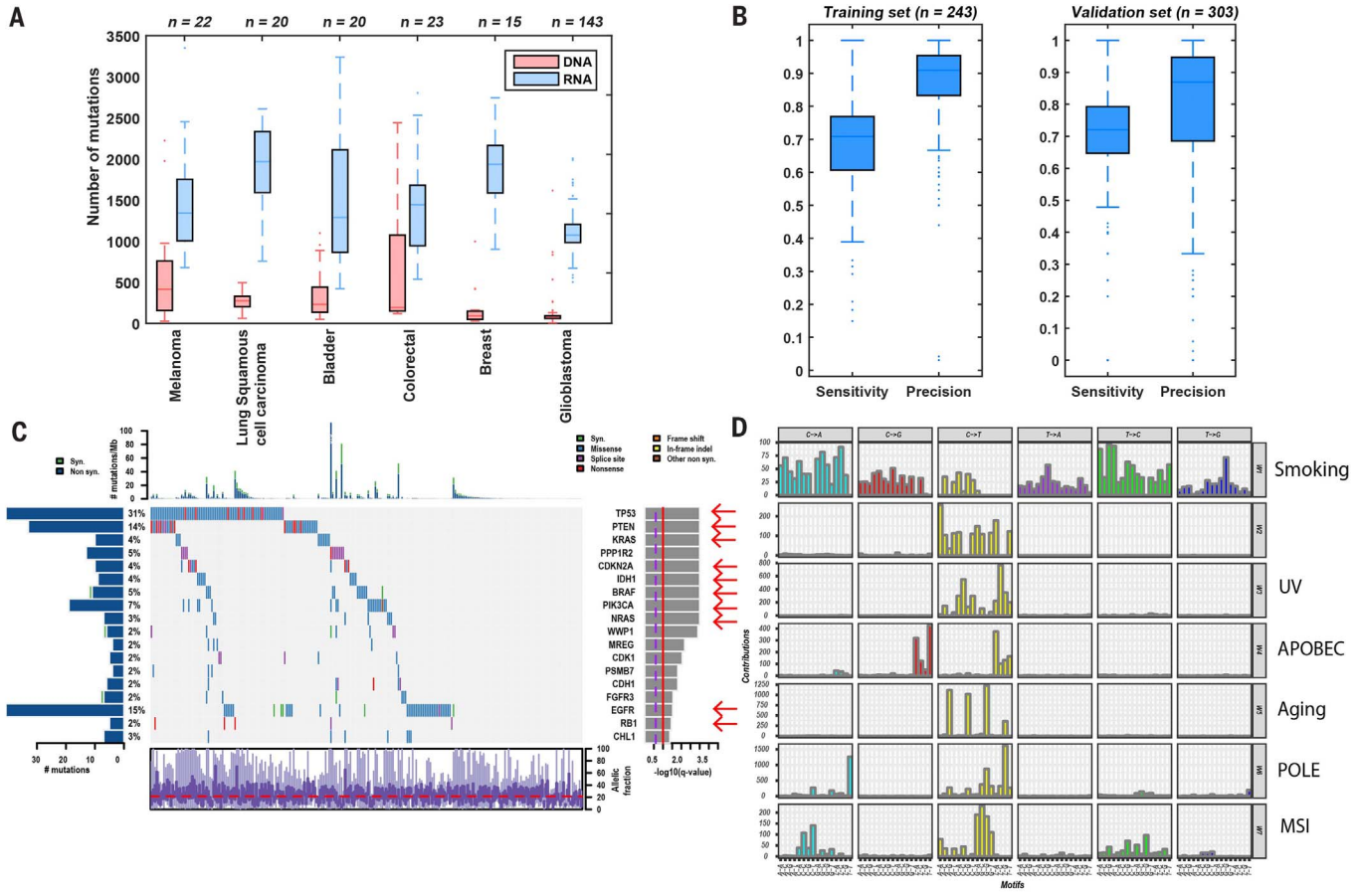


Fig. 1. Validation of RNA-MuTect in TCGA samples. (**A**) Total number of mutations detected before filtering in DNA (red) and RNA (blue) across samples in each TCGA cohort. (**B**) Sensitivity and precision of sufficiently covered sites across training and validation samples. Box plots show median, 25th, and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots. (**C**) Co-mutation plot with mutations across the 243 TCGA samples, overall frequencies, allele fractions, and significance levels of candidate cancer genes ($Q < 0.05$) identified by applying MutSig2CV (*24*) on the mutations detected in the RNA. Genes marked with a red arrow were also identified as significantly mutated in the DNA. (**D**) Mutational signatures identified by SignatureAnalyzer (*25*) on the basis of mutations detected in the RNA. The mutational signatures identified are (i) a mixture of smoking and nucleotide-excision repair signatures (W1, combination of COSMIC signatures 4 and 5, cosine similarities of 0.7 and 0.75, respectively); (ii) UV (W3, COSMIC signature 7, cosine similarity = 0.95); (iii) APOBEC (W4, COSMIC signature 13, cosine similarity = 0.9); (iv) aging (W5, COSMIC signature 1, cosine similarity = 0.9); (v) POLE (W6, COSMIC signature 10, cosine similarity = 0.88); (vi) MSI (W7, COSMIC signature 15, cosine similarity = 0.8); and (vii) W2, a signature found only in the RNA.

mutations as false positives if they were detected in the RNA but not in the DNA while having sufficient coverage in the DNA. Although these mutations could in theory be true RNA-only mutations generated via RNA-specific processes (but not in the RNA-editing databases), it is more likely that they are in fact present in the DNA but at allele fractions too low to be detected, as our detection sensitivity calculations assume that the underlying allele fractions of a mutation are the same in the DNA and RNA. Although these two allele fractions are often close, they can vary as a result of variable gene- and allele-specific expression in different cell types within the sample. One way to test this is by examining the correlation between the number of RNA-only mutations and the number of true positive mutations detected in both RNA and DNA. We observed a high correlation (Spearman $R = 0.6$, $P = 4.2 \times 10^{-30}$). This indicates that many of the RNA-only mutations are likely also in the DNA, because we would not expect any correlation to exist between false positive (generated by either noise or RNA processes) and true positive mutations. Nonetheless, we continued with our conservative approach throughout this study and considered all RNA-only mutations detected in sufficiently covered corresponding DNA loci to be false positive mutations. Overall, we conclude that high-precision analysis of somatic mutations based on RNA is achievable despite the apparent limitations in calling mutations de novo from RNA-seq data, allowing for most cancer-associated genes as well as mutational processes to be revealed from RNA-seq data.

Finally, to evaluate the performance of RNA-MuTect on normal tissues, we applied it to a set of 35 tumor-adjacent normal samples collected in TCGA, wherein DNA and RNA were co-isolated from the same sample (table S1). After ensuring that the samples were not contaminated with tumor cells (*26*), we detected 114 DNA-based mutations, with a median allele fraction of 0.06. These mutations and their low allele fractions reflect the existence of small, yet macroscopic, clones in these samples, as expected in normal tissues. Of only eight mutations detected in the DNA that had sufficient sequencing coverage in the RNA to enable detection (*21*), three were indeed detected, one had evidence in two reads (just below our detection level), and the remaining four had no supporting reads in the RNA (table S2). Similarly, the 175 RNA-based mutations had an average allele fraction of 0.07; of the 86 that were sufficiently covered in the DNA, 13 mutations were detected. Overall, the number of RNA-only mutations per sample was very low (median of 1, average of 2; table S2). Because only half of the RNA-based mutations had sufficient coverage in the DNA, we conservatively estimated

the total number of false positive RNA-based mutations per sample to be between 2 and 4.

Overall, when applying RNA-MuTect to normal samples with coextracted DNA and RNA data, we found that DNA mutations with allele fractions of >0.07 could be detected in the RNA in cases where the gene was sufficiently highly expressed. As a specific example, a mutation with an allele fraction of 0.05 requires coverage by at least 124 reads in order to have >95% chance of being detected, and ~17% of a typical transcriptome from a TCGA RNA-seq sample is covered to that depth (fig. S1B). More important, RNA-MuTect detected a low number of potential false positive calls per sample in normal tissues, consistent with what we found in our cancer samples.

Detecting somatic clonal expansions in normal tissues

After establishing RNA-MuTect's performance on both cancer and normal samples, we sought to study somatic mutations across a comprehensive collection of normal tissues by analyzing RNA-seq data from the GTEx project (*20*). For a mutation to be detected in bulk RNA extracted from a normal tissue, a macroscopic clone that harbors and expresses the somatic mutation needs to contribute a sufficient amount of RNA such that the signal can be observed over the background RNA from other cells in the sample (e.g., muscle and fat cells typically do not proliferate, thus diluting the signal from the expanding clone) (Fig. 2A). Thus, the ability to detect a somatic mutation depends on (i) the clonal diversity of the sample, (ii) the depth of sequencing, and (iii) the expression level of the mutated gene. In the GTEx dataset, RNA was extracted from a relatively large amount of tissue material [~20 mg of tissue, estimated to represent 30,000 to 730,000 cells depending on tissue type (*21*)], limiting our ability to identify mutations present in microscopic clonal populations. We did, however,

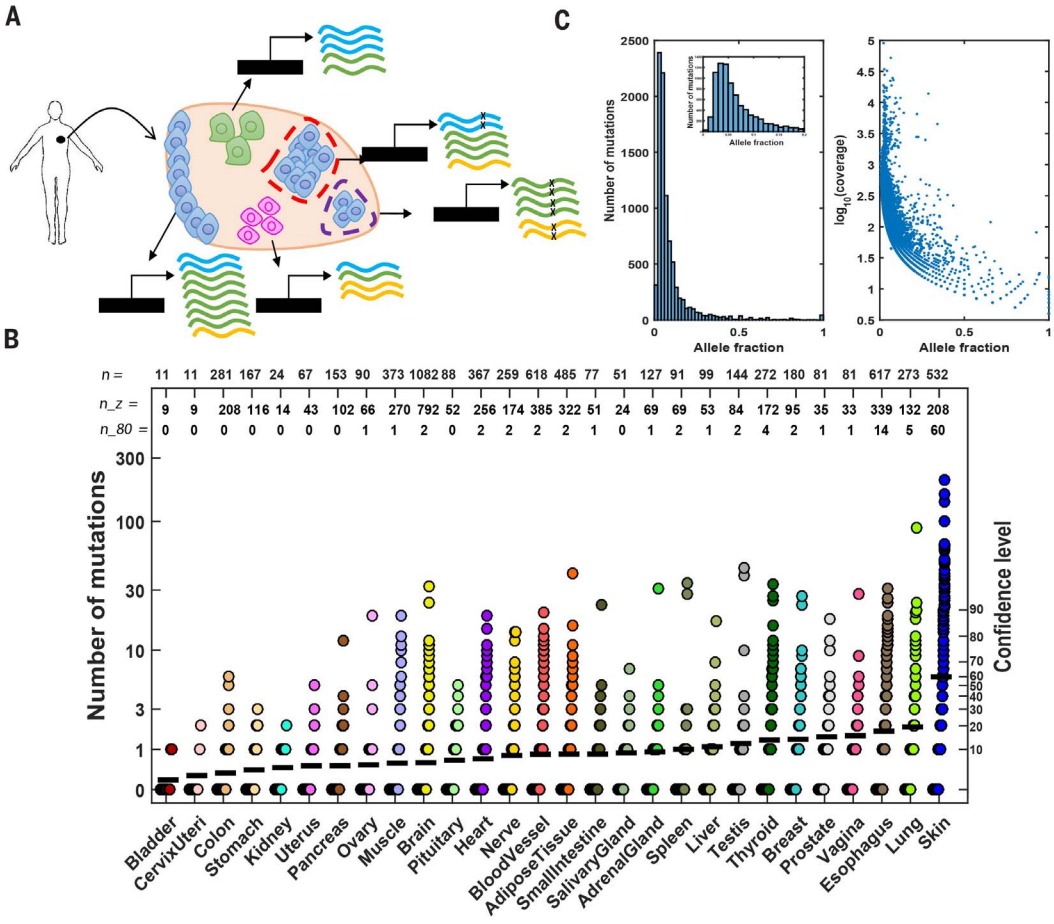


Fig. 2. Somatic clonal expansion in normal tissues. (**A**) An illustration of the composition of bulk RNA extracted from a normal human tissue. The biopsy consists of three different cell types that express different transcripts (marked in blue, green, and yellow) at different levels. Blue cells represent cells with a higher probability to form clones. Two clones, small and large, are denoted by purple- and red-dashed outlines, respectively. Mutated reads are marked with an X. The allele fractions of the mutations in the blue and green genes are the same (0.25; 2/8 and 4/16 reads, respectively), despite the different clone sizes. Additionally, the allele fraction of the mutation in the yellow gene is higher than the allele fractions of the mutations in the blue and green genes (0.33; 2/6 reads), even though the yellow mutation is supported by the same (or smaller) number of reads. These scenarios illustrate the challenge of identifying somatic mutations in bulk normal tissue due to a mixture of cell types and the relatively small clones. Moreover, inferences about clone size are limited because different cell types exist in different proportions and express transcripts at different levels. (**B**) Numbers of mutations detected in RNA-seq of 28 of the 29 studied tissues (we did not detect mutations in six fallopian tube samples). Each sample is represented by a circle. Black horizontal bars represent mean numbers of mutations in each tissue type. A confidence level from our estimation of false positives in the validation data is indicated in the right *y* axis. Specifically, this confidence level is computed as the *x*th percentile on the number of false positive calls (RNA-only mutations in DNA-powered sites) found in the validation set. "*n*" values represent the total number of samples analyzed in each tissue; "*n_z*" values represent the number of samples in which no mutations were detected; and "*n_80*" values represent the number of samples in which more than 13 mutations were found (equivalent to a confidence level of 80%). (**C**) Left: Distribution of allele fraction across all samples in which somatic mutations were detected. Inset: Mutations with allele fraction ≤ 0.2. Right: Allele fraction as a function of $\log_{10}$(coverage) for all detected mutations.

expect to detect macroscopic clones harboring mutations found in ~10% of the cells.

Applying RNA-MuTect to 6707 RNA-seq samples against their matched-blood DNA, which spanned 29 human tissues and 488 individuals (*21*), we detected 8870 somatic mutations in 37% (2519) of the samples, representing nearly all individuals (95%, 467/488; Fig. 2B and table S3). Applying our conservative estimate based on the TCGA data of two to four false positives per sample, 374 samples across 24 tissues had more than four mutations (within these, 106 samples across 13 tissues had more than 13 mutations, which is the conservative estimate at the 80th percentile of false calls). Note that mutations detected in samples with four or fewer mutations are not necessarily false positives; for example, some of these samples harbored known cancer driver mutations that likely increased cell fitness. The analyses described below provide evidence indicating that many of the detected mutations are somatic mutations that reflect clonal expansions in normal tissue.

Similar to what we observed from analyzing the tumor-adjacent normal samples from TCGA, the median allele fraction of the mutations in the GTEx normal tissue samples was 0.05 (Fig. 2C). Although our ability to detect low-allele fraction mutations in both DNA and RNA in GTEx samples was limited because they were extracted from adjacent but different samples, we were able to experimentally validate 5 of 28 mutations by deep sequencing (table S4) (*21*). Consistent with the majority of mutations being passengers, like we observe in cancer, ~59% of the GTEx mutations were missense mutations (fig. S6). However, we also found that a few mutations in normal tissue types matched mutations observed in their corresponding cancer types (table S5). Overall, these results support the idea that macroscopic clonal expansion occurs across many normal tissues throughout the body.

As expected, we found a negative correlation between RNA-seq coverage and allele fraction (Spearman $R = -0.8$, $P < 10^{-200}$; Fig. 2C) due to a higher probability of identifying low-allele fraction mutations in highly covered sites. However, after correcting for detection sensitivity [given the mutation allele fraction and the effective gene coverage (*21*)], we also observed a negative correlation between expression level and expected number of mutations (fig. S7). Similar findings in cancer are attributed to transcription-coupled repair, which is more active in highly expressed genes (*27*–*29*).

The tissues that typically harbor the greatest number of mutations are skin, lung, and esophagus. Associations between cancer incidence in these tissues and environmental factors such as ultraviolet (UV) radiation, air pollution, smoking, and nutritional habits were previously shown (*30*–*36*). Of note, sorting the normal tissues by mutation frequency rather than by the average number of mutations yielded essentially the same order (fig. S8). Looking at tissue subregions, we found that non–sun-exposed skin had more mutations than nonexposed skin and contained the highest number of mutations overall. Similarly, esophagus mucosa, from which esophageal squamous cell carcinomas derive (rather than from either the gastroesophageal junction or esophagus muscularis), had the second-highest mutational burden (fig. S9). Interestingly, the only tissue with a significant difference in the number of mutations between males and females was breast ($P = 2.1 \times 10^{-5}$, two-sided Wilcoxon test; fig. S10), reflecting the observation that breast tissue samples from males in the GTEx dataset are mainly composed of fat cells, whereas female breast tissue also includes epithelial cells.

Finally, we examined whether somatic mutations could be detected in the blood (*21*). Focusing on a previously defined set of 332 single-nucleotide variants detected in the blood of healthy individuals (*3*), we identified 87 mutations in the DNA across 83 individuals (17% of the studied individuals). For each of these 83 individuals, we next tested whether the exact variant was present in other solid tissues from the same individual. Only seven mutations were found in at least one RNA sequencing read in other tissues (each in a different individual) across different tissue types (five in brain, one each in thyroid and heart; table S6). This result most likely suggests that blood had been captured in the tissue samples. Previous results found an increase in the number of detected mutations above the age of 70 (*3*). Although the oldest person in our dataset was 70 years old, we did observe a trend (with borderline significance: $P = 0.049$, one-sided Wilcoxon test) in which these 83 individuals were older than the rest of the cohort.

Clonal expansion increases with age and tissue-specific cell proliferation rate

Several factors can affect the number of mutations accumulated in normal tissues: (i) age, (ii) accumulated DNA damage, and (iii) a tissue's propensity for forming macroscopic clones. All are expected to be more prominent in tissues with a higher cell proliferation rate (*37*, *38*). To test for these associations, we examined whether the age of the individual correlated with the average number of accumulated mutations across tissues. After the age of 45 (the cohort median age), both the number of CpG>T mutations (aging signature) and the total number of mutations significantly increased ($P = 0.001$ and $P = 2.2 \times 10^{-4}$, respectively, one-sided Wilcoxon test; Fig. 3A, top row). This significant association remained after (i) controlling for the number of tissues sequenced in each individual (table S7) and (ii) splitting all individuals into three age groups (fig. S11, A, B, and E). As expected, when considering the top 10 tissues with the highest level of cell proliferation (as determined by *MKI67* expression, a marker of proliferation) (table S8), this relationship became more significant for the total number of mutations ($P = 2.3 \times 10^{-5}$, one-sided Wilcoxon test) and remained similar for the aging mutations ($P = 0.004$, one-sided Wilcoxon test). In the 10 tissues with the lowest cell proliferation, no significant association with age was observed.

Next, we tested whether there was a tissue-specific association with age. A significant association was detected in skin and esophagus tissues ($P = 2.1 \times 10^{-6}$ and $P = 1.5 \times 10^{-5}$, respectively, one-sided Wilcoxon test; Fig. 3A, bottom row). When considering sun-exposed and non–sun-exposed skin separately, we found that although the number of observed mutations increased with age in both skin types, the increase was significantly greater in sun-exposed versus non–sun-exposed skin [odds ratio = 3.29 and 1.26, $P = 1.7 \times 10^{-8}$ and 0.68, respectively (Fisher's exact test), using the number of mutations below or above the tissue's median mutation number (= 2); fig. S11C]. Because these samples derive from the same tissue type, and hence are expected to have a similar cell proliferation rate, this result suggests that either (i) increased exposure to UV light and other environmental factors contributes to DNA damage as an individual ages, or (ii) the size of clones increases in both skin types with increasing age, but the clones in sun-exposed skin enable us to detect the mutations that were acquired earlier in life. Differences were also observed when testing esophagus-derived mucosa, gastroesophageal junction, and muscularis tissues [odds ratio = 4.3, 0.87, and 4.7; Fisher's $P = 2.6 \times 10^{-7}$, 1, and 0.17, respectively (median = 2); fig. S11D]. The lack of association in other tissues could be due to either low cell proliferation rates or the presence of clones below our detection threshold.

We next directly examined whether cell proliferation was associated with the number of accumulated mutations across tissues. Indeed, *MKI67* expression was significantly higher in tissues with a higher number of mutations ($P = 8.2 \times 10^{-4}$ and $P = 1.2 \times 10^{-4}$ for all primary and subregion tissues, respectively, one-sided Wilcoxon test; overall Spearman correlation of $R = 0.44$ and $P = 0.01$; Fig. 3B and fig. S11F) (*21*). Overall, these data suggest that both aging and exposure to mutagenic factors contribute to the number of accumulated mutations, especially in tissues with high cell proliferation rates (*37*, *39*).

Mutational signatures in normal tissues

In addition to the identified aging mutations (CpG>T), we examined whether and which other mutational processes were active in normal samples by applying SignatureAnalyzer (*21*, *25*). Because most samples had a small number of mutations (fig. S12) (*21*), we analyzed only the 169 samples with ≥10 mutations. SignatureAnalyzer identified a UV signature in skin samples. This UV signature is common in melanoma and has been reported in skin fibroblasts and normal skin samples (*17*, *36*). When examining sun-exposed and non–sun-exposed skin separately, the UV signature was active in 62/67 sun-exposed samples and only 1/5 non–sun-exposed samples (Fig. 3C; $P = 5.7 \times 10^{-4}$, Fisher's exact test). Interestingly, all the skin samples with ≥10 mutations analyzed here were from the 447 individuals of European ancestry. In contrast, none of the samples from the 74 individuals of African ancestry had more than six mutations (Fig. 3C), regardless of sun

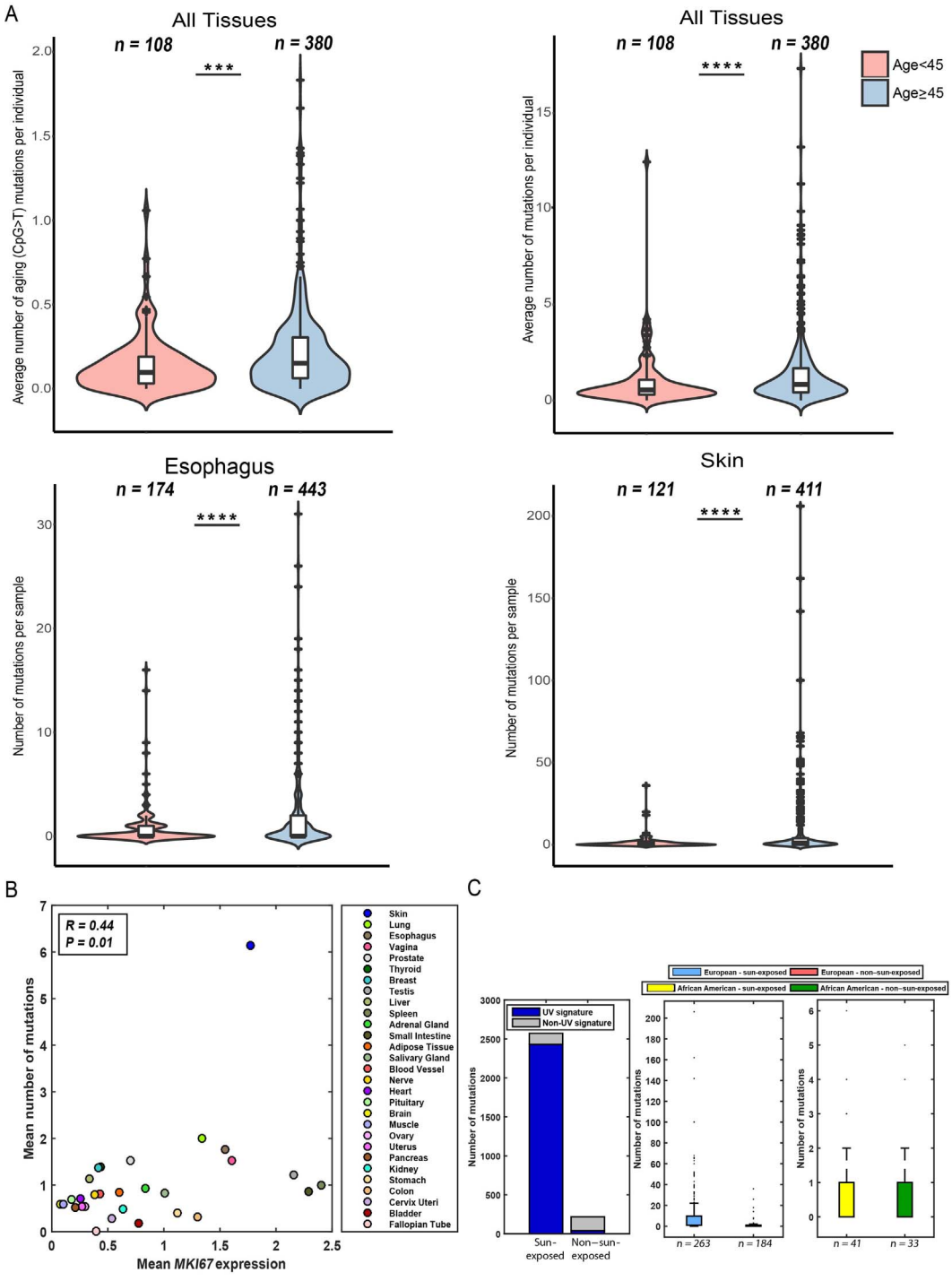


Fig. 3. Mutation load is associated with age and tissue-specific proliferation rate. (**A**) Top: Differences in the average number of aging-related mutations and total number of mutations before and after the age of 45 (left and right, respectively). Bottom: Differences in mutation number in esophagus and skin samples before and after the age of 45 (left and right, respectively). Box plots show median, 25th, and 75th percentiles in each group. Black crosses represent the outliers; asterisks represent significance levels. (**B**) Mean expression of the proliferation marker *MKI67* versus the average number of mutations found in each tissue. (**C**) Left: Number of mutations associated with the UV signature in sun-exposed and non–sun-exposed skin samples. Center: Number of mutations found in sun-exposed and non–sun-exposed skin samples taken from individuals of European ancestry. Right: Number of mutations found in sun-exposed and non–sun-exposed skin samples taken from individuals of African American ancestry. Boxes and whiskers are box plots with dots reflecting outliers.

exposure. Indeed, no difference in mutations was found between sun-exposed and non–sun-exposed skin among African American individuals (an average of 0.87 and 0.81 mutations, respectively; $P = 0.58$, one-sided Wilcoxon test). Overall, skin was the only tissue that showed a significant difference between the total number of mutations detected in European-ancestry versus African-ancestry samples ($P = 1.9 \times 10^{-5}$, one-sided Wilcoxon test; fig. S13).

Mutations in cancer genes in normal tissues

To determine whether somatic mutations in normal tissues occur in known cancer genes, we tested the frequency of nonsynonymous mutations within Cancer Gene Census (CGC) genes (*40*). This CGC set represents genes in which mutations have been causally implicated in cancer. We found that 3% of the samples and 33% of the individuals carried at least one nonsynonymous mutation in a CGC gene. Examining the tissues enriched with nonsynonymous mutations (*21*), we found that skin, esophagus, adipose, adrenal gland, and uterus tissues were significantly enriched with mutations in CGC genes (empirical $Q < 0.1$) after controlling for both gene length and coverage (fig. S14A).

Consistent with previous findings (*17*, *18*), the most frequently mutated cancer genes in our data were *TP53* and *NOTCH1* (Fig. 4A). Examining

whether the number of mutations differed between samples carrying *TP53* mutations and those that did not, we found that the *TP53*-associated samples had significantly more mutations ($P = 9.2 \times 10^{-9}$, two-sided Wilcoxon test). To test whether these *TP53* mutations conferred a growth advantage to the cell, we analyzed their allele fraction level relative to all other detected mutations in the same sample (*21*). Indeed, the allele fractions of *TP53* mutations were significantly higher than other mutations in the corresponding sample (empirical $P < 0.02$; fig. S14B). Similarly, we also found that the *NOTCH1*-mutated cases had a significant increase in the overall number of mutations ($P = 1 \times 10^{-7}$, two-sided Wilcoxon test) as well as a significantly higher allele fraction of the *NOTCH1* mutation (empirical $P < 9.9 \times 10^{-4}$; fig. S14C). These findings were independent of *TP53* and *NOTCH1* expression levels (fig. S14, B and C). This higher allele fraction of the *TP53* and *NOTCH1* mutations relative to other mutations in the same samples suggests that these mutations appear early in the history (i.e., the trunk) of these clones and potentially affected by loss of heterozygosity. However, because early appearance in the trunk does not guarantee that these mutations conferred a growth advantage, we cannot rule out the possibility that these early events are the result of genetic drift; we do consider this possibility unlikely, however, because both *TP53* and *NOTCH1* are known cancer genes. Overall, samples carrying *TP53* or *NOTCH1* mutations were found only in skin and esophagus tissues (with equal proportions in each tissue; table S3), but no samples harbored mutations in both of these genes.

We next examined whether any of the ~1760 recurrently mutated sites (hotspots) in known cancer genes were observed in normal tissues (table S9). We found 30 such mutations in eight tissues that overall included 27 hotspots in 12 genes (Fig. 4A and table S10). The gene with the greatest number of detected hotspot mutations was *TP53* with 16 known hotspot mutations in both skin and esophagus samples, 14 of which were observed once in our dataset (Fig. 4A). In total, 10 of these mutations were previously reported in either (i) normal human skin or peritoneal or uterine lavage fluids taken from healthy women, or (ii) human pluripotent stem cells (*16*, *17*, *41*, *42*). Reviewing the International Agency for Research on Cancer (IARC) TP53 database (*43*), we found that all of these mutations were annotated as deleterious by the SIFT algorithm (*44*). Interestingly, although all of the mutations were annotated as loss-of-function in yeast, three (R248Q, R248W, R282W) were reported to have gain-of-function activities (*45*). R248Q knock-in mice showed an earlier onset of tumor formation and reduced lifespan, as well as an expansion of

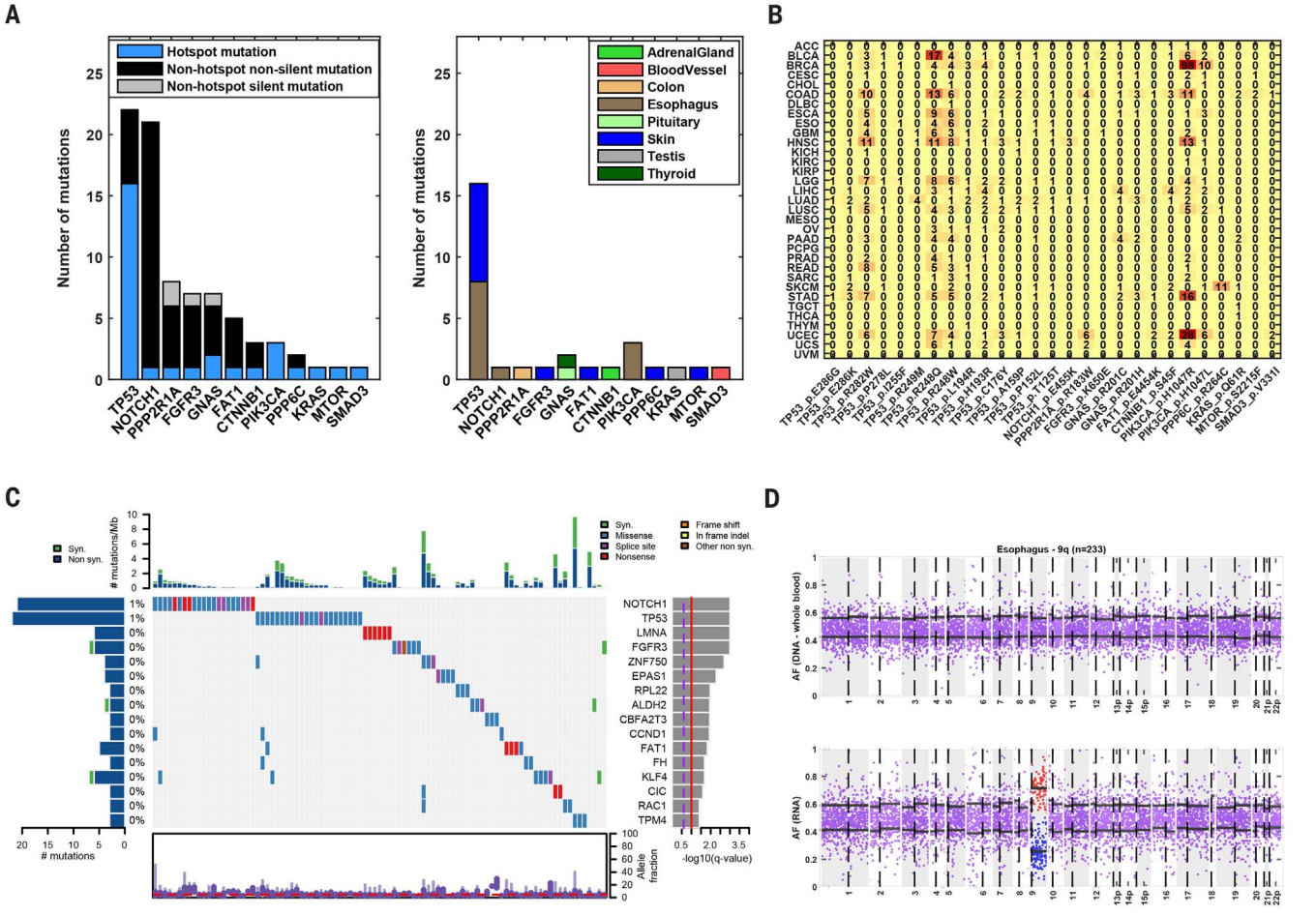


Fig. 4. Mutations in cancer genes across normal tissues. (**A**) Genes in which hotspot mutations were detected. Left: Number of hotspot mutations detected in each gene, and numbers of silent and nonsilent mutations that are not in hotspots. Right: Normal tissues in which the hotspot mutations were detected. All hotspot mutations except two (*FAT1* p.E4454K; *FGFR3* p.K650E) were annotated as pathogenic. (**B**) Occurrences of each hotspot mutation found in different TCGA cohorts. (**C**) Co-mutation plot for genes significantly mutated in a pan-normal analysis, ordered by their significance level (by MutSig2CV); data show 93 of 6707 samples with at least one mutation in these genes and the overall frequency among samples with at least one mutation. The distribution of allele fraction of mutations appears at the bottom. (**D**) Allelic imbalance in chromosome 9q of a normal esophagus sample. Top: Allele fraction of 233 heterozygous sites based on DNA from a matched-blood sample. Bottom: Allele fraction of heterozygous sites based on RNA from the esophagus sample. The black horizonal lines indicate the mean allele fraction per chromosomal arm of sites with allele fraction smaller or greater than 0.5.

hematopoietic and mesenchymal stem cell progenitors (*46*). The R248W variant was involved in multiple gain-of-function activities, including promotion of cell invasion (*47*) and increased cell proliferation (*48*) among others (*45*). The R282W variant increased colony formation (*49*). We found that these 14 hotspot sites shared some tissue specificity with the corresponding primary cancerous tissue, wherein four skin mutations and five esophagus mutations were also observed in melanoma and esophagus TCGA samples, respectively (Fig. 4B).

Among the other 14 non-*TP53* hotspot mutations, all but two were annotated as pathogenic by FATHMM (*50*), and seven were also observed in their corresponding cancer type (Fig. 4B). Three *PIK3CA* mutations in the p.H1047L and p.H1047R hotspots, which are common in multiple cancers (including esophageal cancer), were observed in normal esophagus mucosa samples. The well-known p.Q61R *KRAS* hotspot mutation found in a normal testis sample of a 58-year-old male was also reported in testicular germ cell cancer (*51*). The p.R183W hotspot mutation in the cell growth regulator *PPP2R1A* detected in a normal colon sample here was also detected in colorectal cancer. Although the β isoform (*PPP2R1B*) was discovered as a tumor suppressor in colon cancer cell lines and primary tumors (*52*), the α isoform had also been observed in a cohort of primary colon tumors (*53*). The hotspot mutation p.S45F in *CTNNB1* (β-catenin) found in the normal adrenal gland sample of a 58-year-old female had previously been detected in adrenocortical adenomas; *CTNNB1* is also significantly mutated in adrenocortical tumors (*10*, *54*, *55*) and when mutated deregulates the Wnt/β-catenin pathway. The hotspot mutation p.R264C in the *PPP6C* gene that we detected in normal skin was also observed in melanoma, wherein this gene was found to be significantly mutated (*56*).

To further explore whether clonal expansion observed in normal tissues was in part due to positive selection, we computed the dN/dS (ratio of nonsynonymous to synonymous substitutions) per gene (*57*), taking into account the trinucleotide context and the mutational spectrum (*21*). We found that both CGC genes and cancer genes listed in Lawrence *et al.* (*24*) were enriched with genes exhibiting a higher rate of nonsynonymous mutations (one-sided Wilcoxon $P = 3.4 \times 10^{-4}$ and $P = 9.3 \times 10^{-4}$, respectively). These data suggest that some of these mutations may confer a selective advantage. Of note, these results become insignificant when removing genes identified in skin and esophagus tissues. This finding could be due to the overall low number of mutations detected in the other tissues; alternatively, it may suggest that clones in skin and esophagus tissues undergo positive selection, whereas clones in the other tissues reflect genetic drift.

To more specifically identify which of these cancer genes are significantly mutated, we performed a pan-normal analysis by applying MutSig2CV (*24*) to all 2519 samples in which we detected at least one mutation, restricting the test to 718 known cancer genes (table S11). This analysis yielded 16 significantly mutated genes, with 99 nonsilent mutations spanning 17 tissues, 93 samples, and 82 individuals (Fig. 4C and fig. S14D). In addition to *TP53*, *NOTCH1*, and *FAT1* previously reported as significantly mutated in normal skin (*17*), we also identified other genes such as *RAC1* and *ZNF750*, which are significantly mutated in melanoma and esophagus squamous cell carcinoma, respectively (*9*, *24*). Overall, our results show that cancer genes and hotspots are present in normal tissues, especially in skin and esophagus tissues.

Allelic imbalance in normal tissues

To study other somatic alterations in normal samples, we developed a method for identifying allelic imbalance across chromosome arms using RNA-seq data (*21*), which is similar to previous approaches used for detecting allelic imbalance (*58*, *59*). To test our approach, we applied it to four TCGA samples for which DNA and RNA were coextracted and showed that the vast majority of allelic imbalance events at the chromosomal arm level detected in the RNA were also found in the DNA, and vice versa (fig. S15). In addition, we found a high correlation between the allele fraction of heterozygous sites in the RNA and in the DNA (R range = 0.45 to 0.7, $P < 8 \times 10^{-225}$; fig. S16), which suggests that approaches developed for detecting allelic imbalance in DNA can also work for RNA.

Similar to a recent concurrent study of normal esophagus DNA (*18*), we identified eight esophagus mucosa samples that had an allelic imbalance in 9q (Fig. 4D and fig. S17). Two of the eight samples also had a nonsense or missense mutation in *NOTCH1* (hypergeometric $P = 0.02$), a gene also located on 9q. The allele fraction of these mutations was relatively high (0.22 and 0.12, respectively) and at the top quintile of their corresponding samples. This might suggest that either the wild-type copy of these chromosome arms was lost, or that the mutated copy was gained. Interestingly, 9q loss was more common in esophageal dysplasia than in esophageal squamous cell carcinoma (*60*). Its detection here in nondysplastic lesions suggests that this may be an early event in the development of dysplasia. One additional sample with 9q imbalance was found to carry mutations in both *TP53* and *FAT1*. An allelic imbalance in 22p and a mutation in *NOTCH1* were also identified in an additional esophagus sample (fig. S17). Finally, we identified a testis sample with a strong allelic imbalance in 17p, with no point mutation detected (fig. S17).

Discussion

This study presents a comprehensive overview of somatic clonal expansion in human tissues. Although the use of RNA to detect somatic mutations is limited to expressed genes, we found that RNA analysis can reveal true somatic variations after accounting for both sequencing and alignment noise; moreover, RNA-based analysis can identify both underlying mutational processes and significantly mutated genes. Our approach enabled us to detect thousands of somatic mutations across all human tissues and in almost all tested individuals, including mutations at cancer hotspots and other cancer genes.

Macroscopic clonal expansion was detected in all tissues. However, greater numbers of accumulated mutations were observed in sun-exposed skin, esophagus mucosa, and lung than in other tissues. All three of these tissues are exposed to carcinogenic environmental factors, emphasizing the contribution of extrinsic factors to the mutagenesis process. Indeed, these tissues are also among those carrying the greatest number of somatic mutations in cancer patients (*28*), consistent with the notion that a non-negligible proportion of the mutations observed in cancer accumulate well before disease (*37*). In both skin and esophagus, we observed an association between the number of mutations in normal tissue and age, suggesting a contribution of somatic mosaicism to the aging phenotype (*61*). The lack of association with age in normal lung tissue may be masked as a result of the effects of other factors that are missing in our data, such as smoking or exposure to air pollution.

Beyond these intrinsic and extrinsic factors, the cellular microenvironment and tissue architecture are likely to influence the differences observed among tissues. Studies of different tumor types have shown differences in both the composition of the microenvironment and the transcriptional program active in each tissue (*62*–*69*). In addition, it was previously argued that tissue compartmentalization can affect the rate at which cancer mutations accumulate (*70*). For instance, the arrangement of the intestinal epithelium into crypts and villi is believed to limit the expansion of fitter cells (*71*). Overall, the complex nature of transformation from a normal to a cancer cell within different tissues is a result of the interplay among genetic and epigenetic events, tissue structure, exposure, and the tissue microenvironment. More comprehensive and dedicated data and metadata from various tissues should be collected to further study these relationships.

Relative to studies focusing on microscopic clones (*17*, *18*), we found a significantly lower number of clonal expansions, even though our scale was much larger and not restricted to a specific set of genes. Although this result can be partially explained by our missing mutations in genes with low expression levels, it also suggests that the majority of clones remain microscopic and do not expand to a size that can currently be detected by bulk RNA-seq. In addition, *TP53* and *NOTCH1* were the most mutated genes in our data with a relatively high allele fraction, but their overall frequency was lower than previously observed in microscopic clones. This suggests that mutations in these gene might not be able to drive clonal growth beyond a certain size without additional genetic, epigenetic, or environmental contributions. Furthermore, it should be noted that we identified known driver genes in some clones but not in many others. This observation may suggest that these clones

do not have greater fitness and are the result of genetic drift. In this study, we decided not to draw any conclusions from the distribution of allele fractions on selection beyond our findings for *TP53* and *NOTCH1* because of the nontrivial relation between variant allele fraction in RNA-seq and clone size. Large-scale studies analyzing DNA sequencing data are needed to better distinguish selection versus drift in macroscopic clones in normal tissues.

The overall low rate of cancer-related events in our data (<10%) most likely reflects both our detection sensitivity and the fact that we have analyzed only a single biopsy from each tissue type in each individual. Given previous results from deep sequencing on much smaller tissue biopsies (*17*, *18*), it is reasonable to assume that we would have detected a larger number of somatic mutations across all normal tissues if we analyzed more biopsies from any given tissue type, and if those biopsies were more enriched with epithelial cells. This implies that although these macroscopic clones have expanded to the point of detection, they would remain harmless and may not develop into cancer until—and only if—additional transforming events occur. Also, the detection of hotspots and other mutations in cancer genes across various normal human tissues emphasizes the need for identifying drivers of the disease while considering the nonpathogenic landscape of mutations. Such findings may have an impact on the selection of therapeutic and vaccination targets.

Understanding the earliest genetic events that occur in human tissues may advance our understanding of aging and cancer. Therefore, initiatives such as the Pre-Cancer Genome Atlas (*72*) will substantially aid our ability to detect and treat the disease in its early stages. Because all individuals in this study are deceased, we cannot determine whether the detected clones would have eventually developed into cancer upon acquisition of additional genetic or epigenetic abnormalities. Studying clonal expansion of normal samples longitudinally as they progress from normal tissue to microscopic clones and finally to macroscopic clones will shed light on which of these premalignant lesions has the capacity to transform into cancer; moreover, such a longitudinal study can reveal the required combinations of genetic and/or epigenetic events needed for transformation.

Methods summary

We developed a method, called RNA-MuTect, for identifying somatic mutations from a tissue RNA sample and its matched-normal DNA. RNA-MuTect includes several filtering steps designed for RNA sequences. RNA-MuTect was validated on both cancer and normal samples from TCGA, wherein DNA and RNA were coextracted from the same samples. A power analysis was performed to evaluate the statistical power of observing a mutation, given the mutation allele fraction and sequence coverage at the site. MutSig2CV and SignatureAnalyzer (*24*, *25*) were applied for identifying significantly mutated genes and mutational signatures, respectively. A context-dependent dN/dS analysis was performed for identifying genes with an excessive number of protein-altering mutations. We applied HaplotypeCaller and fitted a beta distribution in order to detect events of allelic imbalance at the chromosome arm level.

REFERENCES AND NOTES

1. C. D. McFarland, K. S. Korolev, G. V. Kryukov, S. R. Sunyaev, L. A. Mirny, Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2910–2915 (2013). doi: 10.1073/pnas.1213968110; pmid: 23388632
2. M. Vermulst *et al*., DNA deletions and clonal mutations drive premature aging in mitochondrial mutator mice. *Nat. Genet.* **40**, 392–394 (2008). doi: 10.1038/ng.95; pmid: 18311139
3. S. Jaiswal *et al*., Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014). doi: 10.1056/NEJMoa1408617; pmid: 25426837
4. A. Poduri, G. D. Evrony, X. Cai, C. A. Walsh, Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 1237758 (2013). doi: 10.1126/science.1237758; pmid: 23828942
5. M. Greaves, C. C. Maley, Clonal evolution in cancer. *Nature* **481**, 306–313 (2012). doi: 10.1038/nature10762; pmid: 22258609
6. Cancer Genome Atlas Research Network, The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015). doi: 10.1016/j.cell.2015.10.025; pmid: 26544944
7. C. W. Brennan *et al*., The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013). doi: 10.1016/j.cell.2013.09.034; pmid: 24120142
8. Cancer Genome Atlas Research Network, Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017). doi: 10.1038/nature20805; pmid: 28052061
9. D.-C. Lin *et al*., Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 467–473 (2014). doi: 10.1038/ng.2935; pmid: 24686850
10. S. Zheng *et al*., Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell* **29**, 723–736 (2016). doi: 10.1016/j.ccell.2016.04.002; pmid: 27165744
11. Cancer Genome Atlas Research Network, Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017). doi: 10.1038/nature21386; pmid: 28112728
12. K. A. Hoadley *et al*., Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018). doi: 10.1016/j.cell.2018.03.022; pmid: 29625048
13. G. Genovese *et al*., Synthetic vulnerabilities of mesenchymal subpopulations in pancreatic cancer. *Nature* **542**, 362–366 (2017). doi: 10.1038/nature21064; pmid: 28178232
14. T. A. Knijnenburg *et al*., Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239–254.e6 (2018). doi: 10.1016/j.celrep.2018.03.076; pmid: 29617664
15. X. Cai *et al*., Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014). doi: 10.1016/j.celrep.2014.07.043; pmid: 25159146
16. J. D. Krimmel *et al*., Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6005–6010 (2016). doi: 10.1073/pnas.1601311113; pmid: 27152024
17. I. Martincorena *et al*., High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015). doi: 10.1126/science.aaa6806; pmid: 25999502
18. I. Martincorena *et al*., Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018). doi: 10.1126/science.aau3879; pmid: 30337457
19. G. Genovese *et al*., Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014). doi: 10.1056/NEJMoa1409405; pmid: 25426838
20. J. Lonsdale *et al*., The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013). doi: 10.1038/ng.2653; pmid: 23715323
21. See supplementary materials.
22. X. Tang *et al*., The eSNV-detect: A computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic Acids Res.* **42**, e172 (2014). doi: 10.1093/nar/gku1005; pmid: 25352556
23. Q. Sheng, S. Zhao, C.-I. Li, Y. Shyr, Y. Guo, Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics* **107**, 163–169 (2016). doi: 10.1016/j.ygeno.2016.03.006; pmid: 27046520
24. M. S. Lawrence *et al*., Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014). doi: 10.1038/nature12912; pmid: 24390350
25. J. Kim *et al*., Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016). doi: 10.1038/ng.3557; pmid: 27111033
26. A. Taylor-Weiner *et al*., DeTiN: Overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018). doi: 10.1038/s41592-018-0036-9; pmid: 29941871
27. M. A. Chapman *et al*., Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011). doi: 10.1038/nature09837; pmid: 21430775
28. M. S. Lawrence *et al*., Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013). doi: 10.1038/nature12213; pmid: 23770567
29. E. D. Pleasance *et al*., A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010). doi: 10.1038/nature08658; pmid: 20016485
30. B. A. Gilchrest, M. S. Eller, A. C. Geller, M. Yaar, The pathogenesis of melanoma induced by ultraviolet radiation. *N. Engl. J. Med.* **340**, 1341–1348 (1999). doi: 10.1056/NEJM199904293401707; pmid: 10219070
31. B. Pesch *et al*., Cigarette smoking and lung cancer—Relative risk estimates for the major histological types from a pooled analysis of case-control studies. *Int. J. Cancer* **131**, 1210–1219 (2012). pmid: 22052329
32. F. Kamangar, W.-H. Chow, C. C. Abnet, S. M. Dawsey, Environmental causes of esophageal cancer. *Gastroenterol. Clin. North Am.* **38**, 27–57 (2009). doi: 10.1016/j.gtc.2009.01.004; pmid: 19327566
33. O. Raaschou-Nielsen *et al*., Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *Lancet Oncol.* **14**, 813–822 (2013). doi: 10.1016/S1470-2045(13)70279-1; pmid: 23849838
34. F. Islami *et al*., High-temperature beverages and foods and esophageal cancer risk—A systematic review. *Int. J. Cancer* **125**, 491–524 (2009). doi: 10.1002/ijc.24445; pmid: 19415743
35. Y. Chen *et al*., Consumption of hot beverages and foods and the risk of esophageal cancer: A meta-analysis of observational studies. *BMC Cancer* **15**, 449 (2015). doi: 10.1186/s12885-015-1185-1; pmid: 26031666
36. N. Saini *et al*., The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. *PLOS Genet.* **12**, e1006385 (2016). doi: 10.1371/journal.pgen.1006385; pmid: 27788131
37. C. Tomasetti, B. Vogelstein, G. Parmigiani, Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1999–2004 (2013). doi: 10.1073/pnas.1221068110 pmid: 23345422
38. C. Tomasetti, B. Vogelstein, Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015). doi: 10.1126/science.1260825; pmid: 25554788
39. L. B. Alexandrov *et al*., Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015). doi: 10.1038/ng.3441; pmid: 26551669
40. S. A. Forbes *et al*., COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015). doi: 10.1093/nar/gku1075; pmid: 25355519
41. N. Nair *et al*., Genomic Analysis of Uterine Lavage Fluid Detects Early Endometrial Cancers and Reveals a Prevalent Landscape of Driver Mutations in Women without Histopathologic Evidence of Cancer: A Prospective Cross-Sectional Study. *PLOS Med.* **13**, e1002206 (2016). doi: 10.1371/journal.pmed.1002206; pmid: 28027320
42. F. T. Merkle *et al*., Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. *Nature* **545**, 229–233 (2017). pmid: 28445466
43. A. Petitjean *et al*., Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. *Hum. Mutat.* **28**, 622–629 (2007). doi: 10.1002/humu.20495; pmid: 17311302
44. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using

the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009). doi: 10.1038/nprot.2009.86; pmid: 19561590
45. P. A. J. Muller, K. H. Vousden, Mutant p53 in cancer: New functions and therapeutic opportunities. *Cancer Cell* **25**, 304–317 (2014). doi: 10.1016/j.ccr.2014.01.021; pmid: 24651012
46. W. Hanel *et al*., Two hot spot mutant p53 mouse models display differential gain of function in tumorigenesis. *Cell Death Differ.* **20**, 898–909 (2013). doi: 10.1038/cdd.2013.17; pmid: 23538418
47. P. A. J. Muller *et al*., Mutant p53 drives invasion by promoting integrin recycling. *Cell* **139**, 1327–1341 (2009). doi: 10.1016/j.cell.2009.11.026; pmid: 20064378
48. W. Yan, X. Chen, Identification of GRO1 as a critical determinant for mutant p53 gain of function. *J. Biol. Chem.* **284**, 12178–12187 (2009). doi: 10.1074/jbc.M900994200; pmid: 19258312
49. M. J. Scian *et al*., Tumor-derived p53 mutants induce oncogenesis by transactivating growth-promoting genes. *Oncogene* **23**, 4430–4443 (2004). doi: 10.1038/sj.onc.1207553; pmid: 15077194
50. H. A. Shihab *et al*., Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013). doi: 10.1002/humu.22225; pmid: 23033316
51. H. Shen *et al*., Integrated Molecular Characterization of Testicular Germ Cell Tumors. *Cell Rep.* **23**, 3392–3406 (2018). doi: 10.1016/j.celrep.2018.05.039; pmid: 29898407
52. S. S. Wang *et al*., Alterations of the PPP2R1B gene in human lung and colon cancer. *Science* **282**, 284–287 (1998). doi: 10.1126/science.282.5387.284; pmid: 9765152
53. Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012). doi: 10.1038/nature11252; pmid: 22810696
54. S. Bonnet *et al*., Wnt/β-catenin pathway activation in adrenocortical adenomas is frequently due to somatic CTNNB1-activating mutations, which are associated with larger and nonsecreting tumors: A study in cortisol-secreting and -nonsecreting tumors. *J. Clin. Endocrinol. Metab.* **96**, E419–E426 (2011). doi: 10.1210/jc.2010-1885; pmid: 21084400
55. L. F. Leal *et al*., Wnt/β-catenin pathway deregulation in childhood adrenocortical tumors. *J. Clin. Endocrinol. Metab.* **96**, 3106–3114 (2011). doi: 10.1210/jc.2011-0363; pmid: 21849527
56. E. Hodis *et al*., A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012). doi: 10.1016/j.cell.2012.06.024; pmid: 22817889
57. M. Kimura, Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977). doi: 10.1038/267275a0; pmid: 865622
58. J. R. González *et al*., A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics* **12**, 166 (2011). doi: 10.1186/1471-2105-12-166; pmid: 21586113
59. U. Weissbein, M. Schachter, D. Egli, N. Benvenisty, Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq. *Nat. Commun.* **7**, 12144 (2016). doi: 10.1038/ncomms12144; pmid: 27385103
60. Z. Z. Shi *et al*., Consistent and differential genetic aberrations between esophageal dysplasia and squamous cell carcinoma detected by array comparative genomic hybridization. *Clin. Cancer Res.* **19**, 5867–5878 (2013). doi: 10.1158/1078-0432.CCR-12-3753; pmid: 24009147
61. R. A. Risques, S. R. Kennedy, Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLOS Genet.* **14**, e1007108 (2018). doi: 10.1371/journal.pgen.1007108; pmid: 29300727
62. E. W. Lin, T. A. Karakasheva, P. D. Hicks, A. J. Bass, A. K. Rustgi, The tumor microenvironment in esophageal cancer. *Oncogene* **35**, 5337–5349 (2016). doi: 10.1038/onc.2016.34; pmid: 26923327
63. X. Sui, L. Lei, L. Chen, T. Xie, X. Li, Inflammatory microenvironment in the initiation and progression of bladder cancer. *Oncotarget* **8**, 93279–93294 (2017). doi: 10.18632/oncotarget.21565; pmid: 29190997
64. J. Villanueva, M. Herlyn, Melanoma and the tumor microenvironment. *Curr. Oncol. Rep.* **10**, 439–446 (2008). doi: 10.1007/s11912-008-0067-y; pmid: 18706274
65. D. F. Quail, J. A. Joyce, The Microenvironmental Landscape of Brain Tumors. *Cancer Cell* **31**, 326–341 (2017). doi: 10.1016/j.ccell.2017.02.009; pmid: 28292436
66. A. E. Place, S. Jin Huh, K. Polyak, The microenvironment in breast cancer progression: Biology and implications for treatment. *Breast Cancer Res.* **13**, 227 (2011). doi: 10.1186/bcr2912; pmid: 22078026
67. I. Soncin *et al*., The tumour microenvironment creates a niche for the self-renewal of tumour-promoting macrophages in colon adenoma. *Nat. Commun.* **9**, 582 (2018). doi: 10.1038/s41467-018-02834-8; pmid: 29422500
68. A. Ghoneum, H. Afify, Z. Salih, M. Kelly, N. Said, Role of tumor microenvironment in ovarian cancer pathobiology. *Oncotarget* **9**, 22832–22849 (2018). doi: 10.18632/oncotarget.25126; pmid: 29854318
69. V. Mittal *et al*., The Microenvironment of Lung Cancer and Therapeutic Implications. *Adv. Exp. Med. Biol.* **890**, 75–110 (2016). doi: 10.1007/978-3-319-24932-2_5; pmid: 26703800
70. J. Cairns, Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975). doi: 10.1038/255197a0
71. H. Quastler, F. G. Sherman, Cell population kinetics in the intestinal epithelium of the mouse. *Exp. Cell Res.* **17**, 420–438 (1959). doi: 10.1016/0014-4827(59)90063-1; pmid: 13672199
72. J. D. Campbell *et al*., The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer Prev. Res.* **9**, 119–124 (2016). doi: 10.1158/1940-6207.CAPR-16-0024; pmid: 26839336

ACKNOWLEDGMENTS

We thank B. Ebert, A. Bass, and T. R. Golub for helpful comments on the manuscript; J. Gastier-Foste and E. Zmuda for helpful information regarding TCGA samples; and M. Miller for help in editing the manuscript. **Funding:** Supported by the Broad-ISF postdoctoral fellowship and the Weizmann award for Women in Science (K.Y.) and by the GTEx LDACC (HHSN268201000029C) and the Paul C. Zamecnick, MD, Chair in Oncology at MGH (G.G.). **Author contributions:** K.Y., P.P., and G.G. conceived the idea; K.Y. and G.G. designed the study; J.H. helped with the MutSig analysis; F.A., J.K., C.S., H.Z., D.L., and D.R. contributed code for the analysis; K.K. and P.A.B. reviewed the pathological samples; J.G. performed the Fluidigm experiments; R.F. generated the docker with the RNA-MuTect pipeline; N.J.H., X.L., E.A.-L., N.S., A.V.S., and K.A. helped with the interpretation of the data; and K.Y. and G.G. wrote the manuscript. **Competing interests:** G.G. receives research funds from IBM and Pharmacyclics. G.G. is an inventor on patent applications related to MuTect, MutSig, and ABSOLUTE. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. The code of RNA-MuTect is available on Zenodo (https://zenodo.org/record/2620062) and in the FireCloud workspace RNA_MuTect (https://portal.firecloud.org). RNA-MuTect is a pipeline that uses multiple tools subject to their own licenses and copyrights. All of these tools are free for academic use. Some tools may require a license from for-profit entities. The details (exact versions) of the different tools that are used in RNA-MuTect are listed in the README.txt in the Zenodo repository of the software.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/364/6444/eaaw0726/suppl/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S15
References (*73*–*88*)

15 November 2018; accepted 2 May 2019
10.1126/science.aaw0726