Published online 15 February 2016

Nucleic Acids Research, 2016, Vol. 44, No. 5 ***2075–2084***
doi: 10.1093/nar/gkw086

The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection

Vinod Kumar Yadav[1], James DeGregori[2,3] and Subhajyoti De[1,3,4,*]

[1]Department of Medicine, University of Colorado School of Medicine, Aurora, CO 80045, USA, [2]Department of Biochemistry and Molecular Genetics. University of Colorado School of Medicine, Aurora, CO 80045, USA, [3]Molecular Oncology Program, University of Colorado Cancer Center, Aurora, CO 80045, USA and [4]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO 80045, USA

Received October 20, 2015; Revised December 18, 2015; Accepted January 31, 2016

ABSTRACT

Mutations acquired during development and aging lead to inter- and intra-tissue genetic variations. Evidence linking such mutations to complex traits and diseases is rising. We detected somatic mutations in protein-coding regions in 140 benign tissue samples representing nine tissue-types (bladder, breast, liver, lung, prostate, stomach, thyroid, head and neck) and paired blood from 70 donors. A total of 80% of the samples had 2–39 mutations detectable at tissue-level resolution. Factors such as age and smoking were associated with increased burden of detectable mutations, and tissues carried signatures of distinct mutagenic processes such as oxidative DNA damage and transcription-coupled repair. Using mutational signatures, we predicted that majority of the mutations in blood originated in hematopoietic stem and early progenitor cells. Missense to silent mutations ratio and the persistence of potentially damaging mutations in expressed genes carried signatures of relaxed purifying selection. Our findings have relevance for etiology, diagnosis and treatment of diseases including cancer.

INTRODUCTION

Recent initiatives have extensively mapped genetic variations in human populations, and also between- and within-sample genetic heterogeneity in cancer. But, in comparison, somatic genetic variation in healthy individuals is much less understood and under-studied (1). Half or more of the point mutations in cancers of self-renewing tissues are suspected to originate prior to tumor initiation (2). Studies by others and us have shown that detectable clonal mosaicism and somatic mutations in non-malignant tissues are linked to cancer risk later in life and survival (3–5). In many cases, somatic mutations are present in clonally expanded cell populations in non-malignant tissues, detectable at tissue-level resolution. For instance, using targeted sequencing, Martincorena *et al*. detected high burden of somatic mutations in clonally expanded cell populations in benign skin biopsy samples (6). Other studies have also reported somatic mutations in non-malignant tissues at a frequency higher than that previously suspected (3,4,7–10). Evidence for roles of somatic mutations in complex traits and diseases other than cancer is also rising (1,11,12). Moreover, intra-tissue genetic variation can be an issue for patient-derived induced pluripotent stem (iPS) cells, since cells from the same donor may not have identical genomes (8). These findings underscore the importance of an assessment of the landscape of somatic mutations in non-malignant tissues, and associated mutation signatures.

MATERIALS AND METHODS

Data acquisition

We obtained exome-sequence data for 140 benign tissue samples from 70 donors with solid tumor from the TCGA (13). The benign samples represented nine solid tissues (bladder, breast, head and neck, liver, lung, prostate, stomach and thyroid), and corresponding blood. We also had tumor exome sequencing data from these patients. Using sequencing data from tumor, benign solid tissue, and blood we identified benign solid tissue and blood specific somatic mutations (Figure 1A). For head and neck squamous cell carcinoma cohort, tumor proximal non-malignant tissue was considered as matched *benign* solid tissue. Any variants arising from tumor DNA contaminations were also excluded after comparing the paired tumor genomes. Average sequencing depth for samples analyzed is 48× and in general depth for tissue and corresponding blood is comparable. Some tissue samples are sequenced at relatively higher (e.g. ∼70× for breast and stomach samples) or lower depth

*To whom correspondence should be addressed. Tel: +1 303 724 6461; Fax: +1 303 724 1799; Email: subhajyoti.de@ucdenver.edu

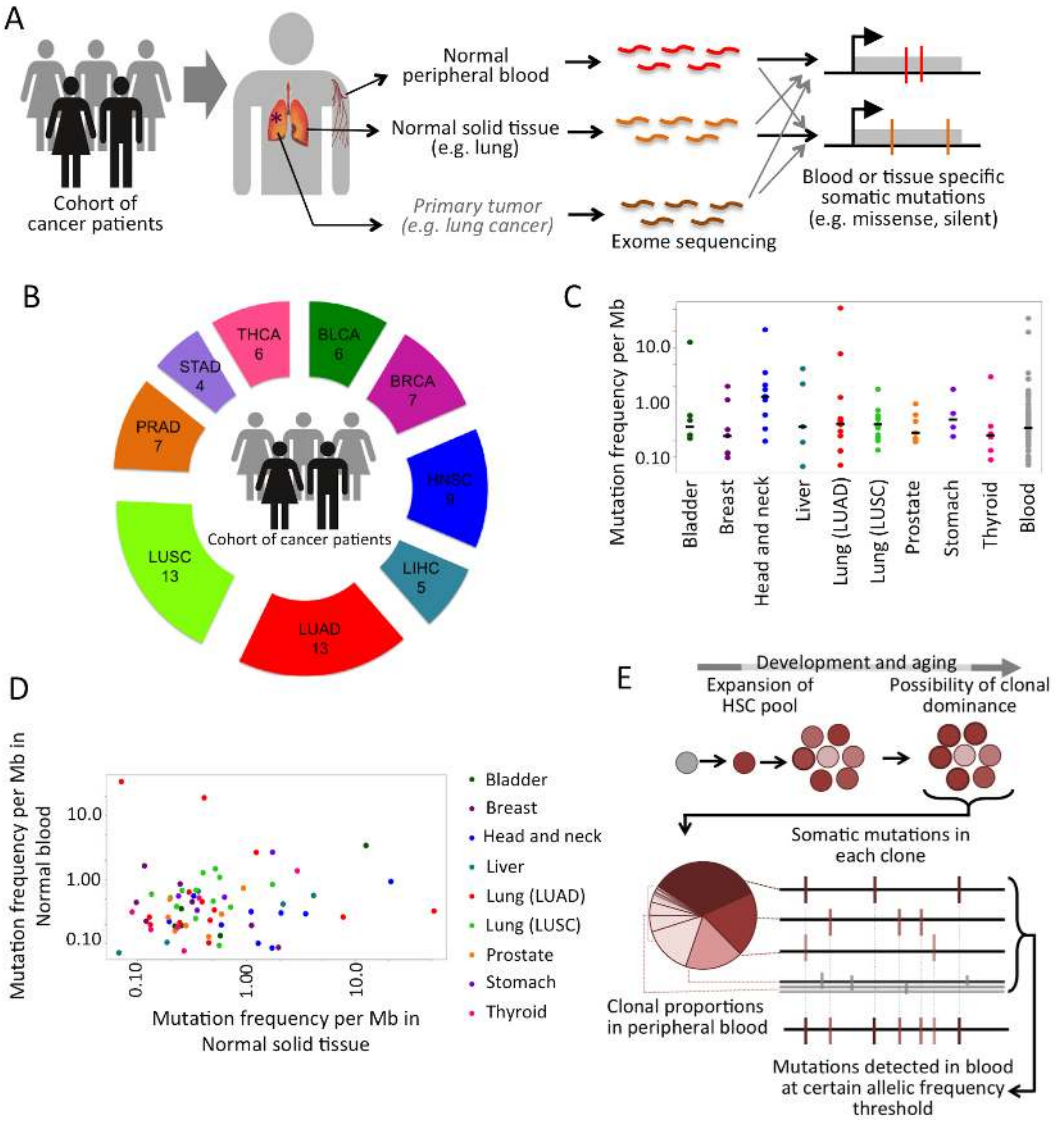

Figure 1. (**A**) A schematic representation showing the pipeline for detecting somatic mutations in apparently benign tissues. (**B**) The TCGA cohorts and the number of donors in each cohort are shown. (**C**) Adjusted mutation detection rate per Mb in exonic regions is shown for different cancer type. For each cancer type, the median value is shown using a black line. (**D**) Scatterplot showing ADMB (adjusted detectable mutation burden) in the matched blood and solid tissues for the donors. (**E**) A schematic representation describing that mutations detected in blood at certain allele frequency threshold can come from one or more major clones.

(e.g. ∼23× bladder and liver samples; Supplementary Table S1). Age, gender and smoking status data for the donors were obtained from the TCGA (13).

Identification and assessment of somatic mutation

To identify high confidence somatic mutations, we used VarScan2 (14) with the following filters: (i) ≥25× coverage at the somatic variant site in both samples (solid tissue and blood), (ii) high-confidence variant call, allele frequency ≥0.1 and *P*-value cut-off 0.05, (ii) variant allele frequency = 0 in the reference sample (i.e. solid tissue for blood specific mutation analysis and *vice versa*), (iii) somatic mutation detection *P*-value > 0.05 cut-off and (iv) variants present in the dbSNP (version 138) database are filtered. Four outliers samples had excessive potential somatic mutations (≥5-fold mutation rate per Mb compared to average mutation rate per Mb in particular tissue type) (TCGA-55-7576:blood, TCGA-44-7661:blood, TCGA-GD-A2C5:bladder, TCGA-44-2665:lung). The excessive mutation burden in the outliers appeared to arise due to a combination of factors including presence of tumor DNA, history of other malignancy and treatment and technical issues related to sample preparation and sequencing (e.g. depth of sequencing). Outliers were excluded from further analysis.

We took multiple measures to estimate the confidence in the reported somatic variant calls, after accounting for possible sources of errors due to misclassified germ line heterozygous SNPs, contamination due to tumor-derived DNA and sequencing error (Supplementary Text and Supplementary Figure S1). Additionally, for a subset of the

samples, matched RNA-Seq data from the same tissues were available. A considerable proportion of the mutant alleles were also detectable in the RNA-Seq data, indicating expression of these variants in the affected tissues. Clonal analysis was performed using approaches used elsewhere (15). In brief, we generated a histogram of allele frequency distribution for a given sample. Using mixtool package (16), we fit mixtures of Gaussian distributions to the data using expectation maximization approach, and decided the number of components in the mixture using Bayesian Information Criterion.

Integrative analysis

We obtained the list of genes associated with DNA repair from the catalog maintained by Richard Wood's group (17). We also obtained the catalog of genes involved in complex traits, and cancer from the NHGRI GWAS catalog (18), and COSMIC (19), respectively. While some of the somatic mutations were in the MHC (major histocompatibility) locus, a vast majority of the mutations (>95%) were elsewhere in the genome. Potential deleterious consequences of the missense mutations were predicted using SIFT (sorting intolerant irom tolerant) algorithm (20). Mas5 condensed expression data (GNF1H) for 33 689 probes from 72 benign human tissues (including multiple blood cell types) were obtained from GNF SymAtlas (21). Probes were mapped onto 33 495 transcripts from 17 185 human genes. The definition of tissue-specific genes was as previously used. We used the ConsensusPathDB (22) for functional analyses, and Cytoscape for representation. Functional consequences of the missense mutations were predicted using SIFT (20), which uses multiple criteria such as amino acid substitution and evolutionary conservation to classify the mutations. Mutations deemed *damaging* are more likely to perturb structure and/or function of the protein, or lead to gain-of-function, while those flagged as *benign* are less likely to do so. The transcription-coupled repair (TCR) score is calculated as the product of (i) the difference in the proportion of C:G > A:T substitution between the highly (top 25 percentile) and lowly expressed (bottom 25 percentile) genes and (ii) transcriptional strand bias in the abundance of C > A versus G > T substitution between the same group of genes, in a given tissue. Tissue enrichment was calculated using geneSetTest of R limma package.

Mathematical modeling

We adopted a linear birth-immigration-death model (23), and considered the following parameter values: Hematopoietic stem cell (HSC) replication rate (λ): once 25–50 weeks, differentiation rate (ν): 0.71L, HSC death rate (α): 0.14L, the rate at which short-term reconstituting cell-derived clones exhaust: 1 per 6.7 weeks, mutation rate: 10^{-8} per base per cell division, based on published studies (24). We calculated the survival probability of an HSC clone as $1 - \mu_0(l)$, where

$$p_0(t) = \left[\alpha + \nu - (\alpha + \nu)e^{(-\lambda+\alpha+\nu)t}\right] \Big/ \left[\lambda - (\alpha + \nu)e^{(-\lambda+\alpha+\nu)t}\right].$$

We changed HSC replication, death and differentiation rates by different proportions, and recalculated the survival probabilities of the clone.

RESULTS

We developed a pipeline to detect somatic single nucleotide mutations in exome-sequencing data (Figure 1A), and used that to analyze 140 samples representing 9 non-malignant solid tissue types (bladder, breast, head and neck, liver, lung, prostate, stomach and thyroid) and paired benign blood samples from 70 donors (cancer patients) to detect exonic somatic mutations (Figure 1B; Supplementary Table S1). Their tumors, as well as matched benign solid tissues and peripheral blood were processed and sequenced by The Cancer Genome Atlas initiative (13). We compared exome bam-files for non-malignant blood, matched solid tissues and also paired tumor samples for each donor, and applied multiple filters to identify somatic mutations specific to non-malignant tissues (allele frequency >0.1; see 'Materials and Methods' section, Supplementary Text and Supplementary Figures S1 and S2 for details). These mutations could be of early developmental origin, or reached high allele frequency selection, drift and/or differentiation bias (Supplementary Text). Variants present in both blood and paired solid tissue from the same donor were not considered, and therefore we might have missed those mutations that occurred very early, before tissue differentiation during embryogenesis. Any potential false positives due to tumor DNA contaminations were also excluded after comparing the paired tumor genomes.

A total of 80% of the samples had 2–39 exonic mutations (median: 8; Figure 1C and Supplementary Figure S3). A substantial fraction of the somatic variants were also detectable in the RNA-Seq data from the same samples, indicating that some of these variants were also expressed (Supplementary Figure S2). The burden of somatic mutations detected in non-malignant tissues in our study appears to be consistent with that reported elsewhere. For instance, Welch *et al*. (25) detected 5–15 exonic mutations per human HSC, and whole genome sequencing of peripheral blood of a super-centenarian (15) detected hundreds of somatic mutations at 60× sequencing coverage. Investigating blood exome-seq data for 2728 cancer patients (with solid tumor), and focusing on 558 cancer associated genes, Xie *et al*. identified 77 somatic mutations (rare truncating variants and known hotspot variants) in 58 individuals (26).

Prevalence of somatic mutations in benign human tissues

Absolute somatic mutation burden could not be directly compared between samples due to technical issues. Therefore, for each sample, we calculated the rate of detectable somatic mutation per base pair, dubbed adjusted detectable mutation burden (ADMB), after considering only the positions covered at sufficient sequencing depth (≥25×). ADMB depends not only on the tissue-specific progenitor cell mutation rates, but also other factors including cell division rate, the actions of selection and drift, clonal make up, and tissue turn over. First, focusing on blood samples, we found that ADMB usually varied within a narrow range between individuals (median: 0.34 per Mb, interquartile range: 0.42 per Mb; Figure 1C). When the samples were grouped according to the original TCGA cohorts, no batch effect was observed (Supplementary Fig-

ure S4). Cross-tissue analysis indicated that ADMB was roughly comparable between the tissue types (median:0.36, inter-quartile range: 0.65 per Mb), although certain tissues such as head and neck had significantly higher ADMB than other tissues (Mann–Whitney U test, *P*-value: 1.66E-02). ADMB in paired blood and solid tissues for the same donor showed no significant correlation (Spearman *r*: 0.152, *P*-value: 2.09E-01, Figure 1D), which was even more apparent when adjusted for age (partial correlation, spearman *r*: 0.109, *P*-value: 3.68E-01). This provided another indication that systematic biases due to batch effects were probably not of concern in our analysis.

Since blood and other tissues are polyclonal, and the constituting clones can contribute to differentiated cells at unequal proportions (e.g. in blood (27,28)), we were able to detect only the mutations present in dominant, genetically distinct clones in the most prevalent cell types in the tissue. Thus, our assessments probably reflect the lower bound of the number of somatic mutations in human tissues (Figure 1E). Nevertheless, note that our tissue-level estimates were consistent with that based on clones derived from single stem cells (25) (Supplementary Figure S5). The majority of the samples did not have sufficient numbers of high frequency somatic mutations to identify the number and proportion of major clones. Nevertheless, high allele frequency somatic mutations probably indicated (i) early developmental origin, (ii) clonal dominance (cell competition, drift and population bottlenecks in the tissue stem cell pool can affect clonal balance and a few genetically distinct clones can become dominant) (29) and/or (ii) skewed clonal contribution in differentiated cells (e.g. actively cycling HSCs contribute disproportionally to peripheral blood) (30), help somatic mutations to reach high allele frequency ('Materials and Methods' section and Supplementary Text).

Factors associated with high rates of somatic mutations in benign tissues

First, integrating age data and grouping blood samples according to patient age, we found that ADMB systematically increased with age (Figure 2A, Spearman *r*: 0.28, *P*-value: 1.82E-02), such that ADMB for the patients in age group of 80–90 years is nearly double of that in the age group of 30–40 years. This is consistent with single clone-based estimates (25) and those derived indirectly from tumor genomes (2). Our result indicates that age-dependent increase in the mutation burden is not only evident in individual clones (2,25), but leaves its signature in terms of high allele frequency somatic mutations in the tissue bulk. Different tissue age differently compared to the age of the individual (dubbed age acceleration), based on their CpG methylation status, which can serve as tissue-specific epigenetic clocks. Using CpG methylation-based tissue-specific age acceleration data (31), we found that the benign tissues with high age acceleration (e.g. breast) had relatively lower ADMB, indicating lighter burden of high allele frequency somatic mutations (Figure 2B). This is consistent with the report that tumor tissues with high age acceleration exhibit fewer somatic mutations in cancer (31), but further work needs to be done to establish causality.

The ADMB in benign lung tissues was significantly correlated with the number of cigarette packs smoked per year, even after adjusting for patient age (partial correlation *P*-value: 5.10E-04; Figure 2C). Smoking leads to oxidative DNA damage and increased burden of acquired mutations in lung cells, but it was interesting to observe that the oxidative DNA damage signature was observed even at the level of high allele frequency somatic mutations in lung. This finding also suggests that not only the mutations that arose during early fetal development, but also those acquired later in adulthood (e.g. due to smoking) could reach dominance in the tissue bulk within a few years or decades, perhaps by cell competition and/or drift. The samples that carried missense mutations in DNA repair genes (17), had significantly higher burden of somatic mutations compared to others (Figure 2D; Mann–Whitney U test, *P*-value: 2.00E-03). We could not validate the DNA repair gene mutations, and also acknowledge that not all missense mutations impair DNA repair function; therefore we interpret these results with caution. Nevertheless, taken together, it is evident that intrinsic and extrinsic factors (e.g. age, smoking) are associated with increased burden of high allelic frequency somatic mutations in apparently benign human tissues.

Substitution patterns and mutational signatures

Analyzing nucleotide substitution pattern across tissue types, we found that C:G > A:T transversions, which are known marker for oxidative DNA damage, were predominant in head and neck, lung (LUSC), stomach and thyroid tissues (Figure 3A). The C:G > A:T substitution signature was unlikely to be artifacts due to sample preparation or sequencing error (Supplementary Text and Supplementary Figure S1). Oxidative DNA damage in tissues such as stomach is well established (32), but the observed proportions of different substitution classes in stomach and thyroid samples could be due to small sample size. The difference in substitution patterns between the two lung cohorts could be due to difference in tissue organization between proximal (LUSC) and distal (LUAD) lung tissues. It is likely that smoking and other mutagenic factors also contributed to the observed substitution patterns. Interestingly, the somatic mutations with relatively low allele frequency had proportionally more C:G > A:T substitutions compared to those present at higher allele frequency (Figure 3B). Furthermore, the proportion of C:G > A:T correlated with age (Figure 3C, spearman $r = 0.22$). We also found that current and former smokers had more C:G > A:T substitution in lung compared to non-smokers (Figure 3D), which is similar to that previously reported (33). In fact, C:G > A:T is a prominent signature of smoking associated oxidative DNA damage in lung (34) and bladder cancer (34). We did not have sufficient smoking data for other cohorts, but it is likely that not only smoking, but also other factors contributed to an excess of C:G > A:T transversions in lung and other organs.

We then examined whether TCR shaped the somatic mutational landscape. We ranked the genes using expression data for CD34+ HSCs (21), and flagged the top 25% of known genes as highly expressed and bottom 25% as lowly expressed. We found that the lowly expressed genes have

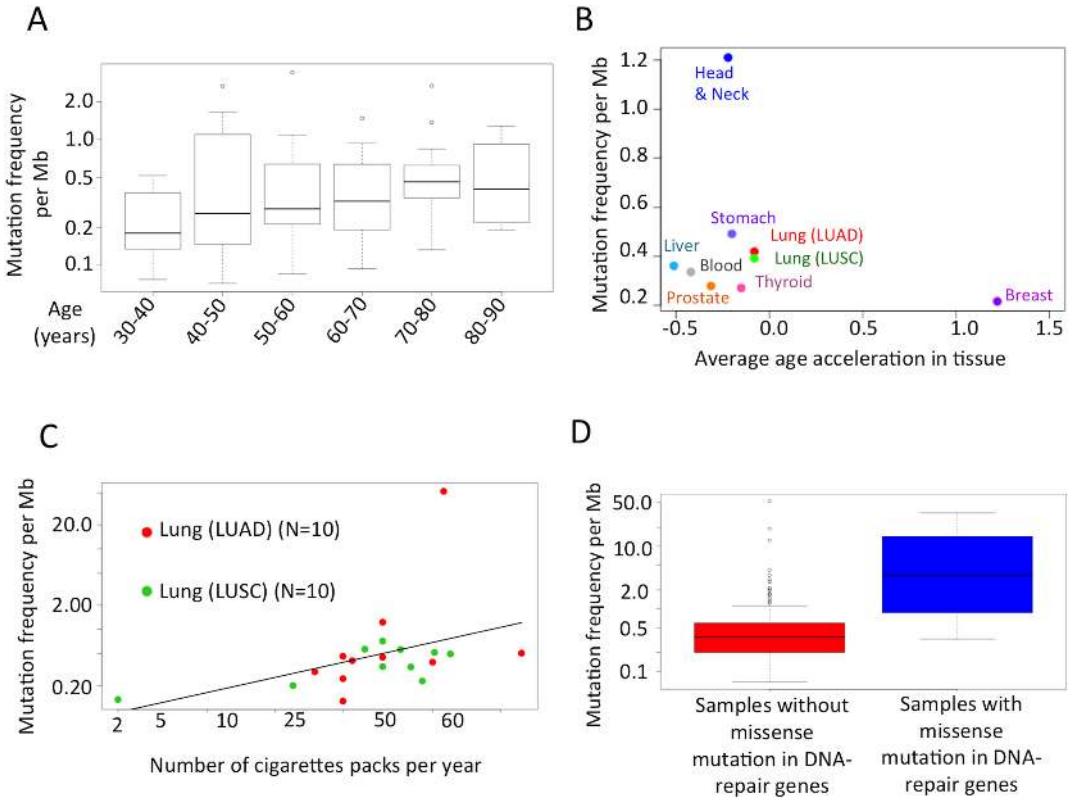

Figure 2. (**A**) Boxplot showing that ADMB in blood increases with age of the donors. (**B**) Scatter-plot showing median ADMB in different tissue types against their average age acceleration score derived from methylation status of a set of informative CpG sites. (**C**) Scatterplot showing association between the numbers of cigarette packs smoked per year and ADMB in lung for the donors. (**D**) Samples with missense mutations in DNA repair genes had significantly higher ADMB compared to other samples in the study.

proportionally more C:G > A:T substitutions indicating higher burden of oxidative DNA damage compared to that in the highly expressed genes (Figure 3E). Furthermore, transcribed strands had more C > A substitution relative to G > T substitution indicating transcriptional strand bias (Figure 3F). We found similar results using expression data for whole blood (Figure 3E–F; Fisher's exact test; *P*-value < 0.05 for all cases). The results were consistent even when alternative expression thresholds were used (Supplementary Figure S6), providing novel evidence for a role of TCR in the somatic mutations in benign tissues.

Likely cell of origin of the detected mutations

Somatic mutations arise during development and aging. Those that arose during early embryonic development, before tissue differentiation stage should be present in both blood and other tissues, and hence cannot be detected in our analysis. On the other hand those, which arose at the terminal stages of differentiation (e.g. lymphoid or myeloid differentiation in blood) are unlikely to reach high allele frequency (e.g. >0.1). Therefore, the majority of the somatic mutations observed in our analysis probably arose in hematopoietic stem and early progenitor cells. We examined whether TCR signature could independently predict the cell of origin of the mutations. We calculated an intuitive index, dubbed TCR score (transcription coupled repair score), that combines both (i) the difference in proportion of C:G > A:T substitution, and also (ii) transcriptional strand bias in C > A versus G > T substitution between highly and lowly expressed genes (see 'Materials and Methods' section for details). We repeated the analysis in Figure 3E–F, using 70 different benign cell types. The score was significantly higher in blood cell-types compared to other cell types (Figure 3G; *P*-value: 5.30E-03); even within blood, lymph node, CD33+ myeloid progenitor cells and CD34+ HSCs ranked at the top of the list (differentiated blood cells had lower ranks; Supplementary Table S2). Other organs (e.g. heart), which had some representation near the top of the list, did not have statistically significant enrichment. Taken together, our observations further supported the hypothesis that majority of the high allele frequency somatic mutations detected in blood were acquired in the HSCs and early lymphoid or myeloid progenitor cell populations.

Missense mutations and pathways

Two-third of the somatic mutations in protein-coding regions in our dataset were missense (Figure 4A), which is consistent with other reports (9,35). In blood, 43 missense mutations were in highly expressed genes (top 25 percentile) and were also deemed as potentially *damaging* by SIFT (20). The affected genes were moderately enriched for signaling and development (Figure 4B). None of the mutations was causally implicated in blood-related diseases or leukemia, and sparse clinical data was insufficient for survival analy-

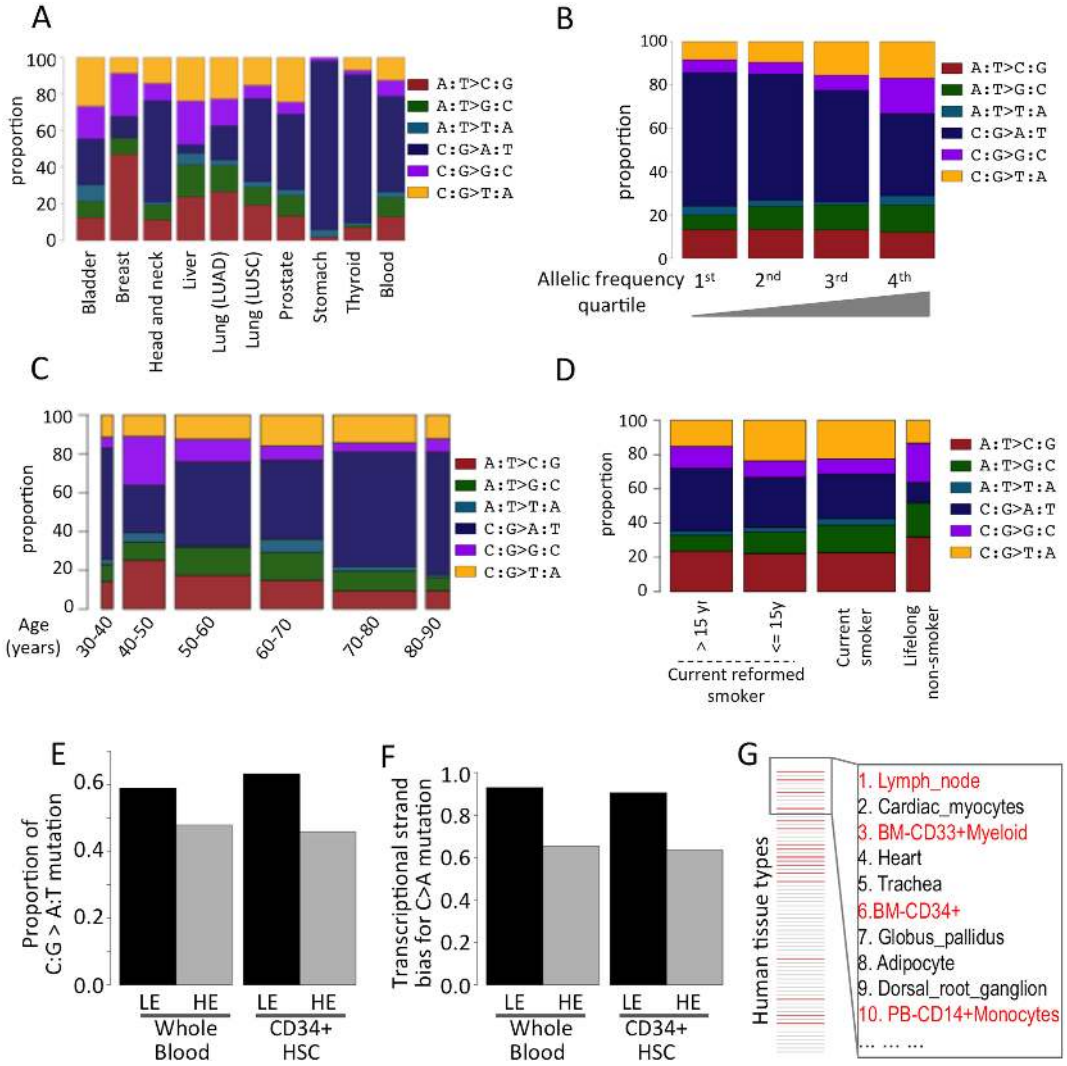

Figure 3. (**A**) Bar-plot showing nucleotide substitution patterns in different tissue types. (**B**) Proportion of different substitution classes for somatic mutations grouped by allele frequency. (**C**) Bar-plot showing age-associated changes in proportion of different types of nucleotide substitutions in blood. (**D**) Bar-plot showing changes in proportion of different types of nucleotide substitutions in lung as a function of smoking habits. (**E**) Proportion of C:G > A:T substitutions in lowly and highly expressed genes, based on expression data in whole blood and CD34+ hematopoietic stem cells. (**F**) Transcriptional strand bias, as estimated by the proportion of G > T/C > A substitution on the transcribed strand for lowly and highly expressed genes. Highly expressed genes have proportionally more C > A substitution than G > T substitution on the transcribed strand than lowly expressed genes. (**G**) Tissues were ranked based on the TCR score, and the top 10 tissue types are shown (blood-related tissues are shown in red).

sis. We repeated the analysis with the genes that were in the top 50 percentile in terms of expression (instead of top 25 percentile) and had potentially *damaging* mutations (Supplementary Figure S7), and detected enrichment for several additional pathway, some of which were associated with immune function and tumorigenesis (*Regulation of Androgen receptor activity*, *P*-value: 2.9E-03*; Chromatin remodeling by SWI/SNF ATP-dependent complexes*, *P*-value: 3.6E-03; *Signaling events mediated by HDAC Class I*, *P*-value: 3.8E-03; *Downstream signaling in naive CD8+ T cells*, *P*-value: 6.6E-03; *Pathways involve in Glioma and Prostate Cancer*, *P*-value: 5.8E-03 and 7.5E-03, respectively). We repeated the functional enrichment analysis, after adjusting for gene expression in blood, and found consistent results (*P*-value < 0.05 in all cases). Interestingly, 12/16 of the missense mutations in these functional categories were potentially *damaging*, higher than that expected based on all the genes expressed in blood (∼1:1, Supplementary Figure S8). The enrichment was apparent even after adjusting for the number of synonymous mutations in these genes, indicating that it was unlikely due to increased local mutation rate. We note that the number of missense mutations in any given individual was small (and hence not suitable for per-individual rigorous pathway enrichment analysis), not all potentially *damaging* mutations affect gene function, and these mutations were not likely to lead to severe functional defects in apparently benign blood. Even then, one might argue that the collective impact of the somatic mutations could potentially contribute to subtle perturbations in cellular processes, and if such mutations occur in clonally expanded cell populations, whether by mutagenesis or selection, the con-

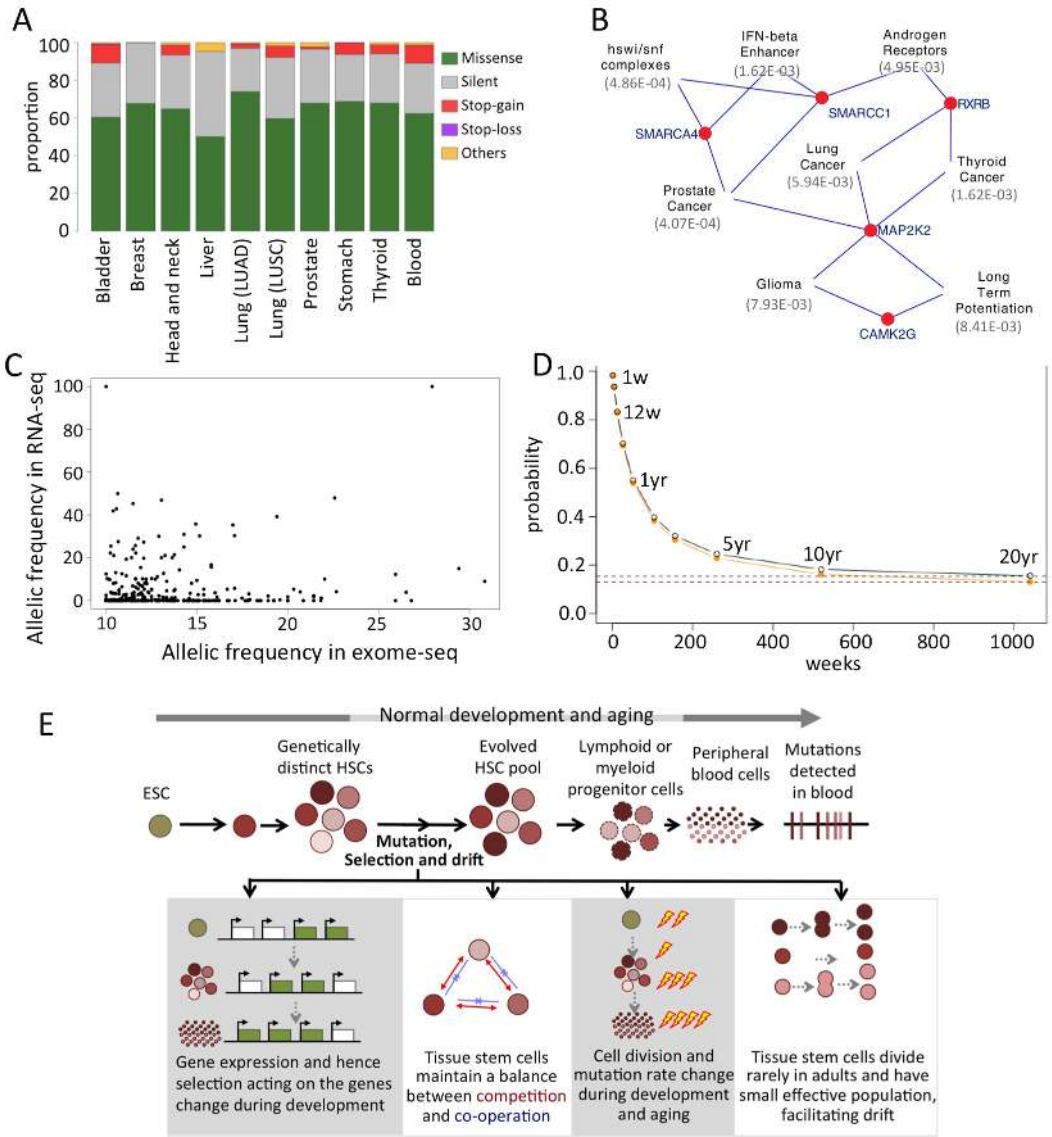

Figure 4. (**A**) Amino acid substitution patterns in different tissue types. (**B**) Pathways that are enriched for potentially *deleterious* mutations in genes that also have high expression in blood. (**C**) Scatterplot showing allele frequency estimates of somatic mutations in RNA and exome-seq data from the same blood samples. (**D**) Graph showing the probability of survival of an HSC clone as a function of time. Black and orange curves show the probability for wild-type HSCs, and those whose death rate is 20% higher, respectively. (**E**) A schematic representation showing effects of mutagenesis, selection, and drift during development and aging on the mutation landscape of non-malignant somatic tissues.

sequences could be detectable at the tissue-level for the individual, as recently reported (5).

Conventional proxies of selection in somatic tissues

Next we analyzed the dataset for classical signatures of selection. First, integrating expression data from the same tissue samples, we found that a majority of the missense mutations were also present in RNA-Seq data from the same samples (Figure 4C). We found consistent results when only potentially *damaging* missense mutations were considered (Supplementary Figure S9) indicating that mutant alleles at the site of missense mutations, even those deemed potentially *damaging* were not necessarily selectively suppressed at the expression level. Second, we found no difference in allelic frequencies between missense, nonsense, and silent mutations in blood (Supplementary Figure S10), and the burden of missense mutations was high (especially relative to silent mutations). We found similar results when considering only the mutations that were in the genes expressed in blood. Furthermore, even when we considered only missense mutations, the ratio of potentially *damaging* versus benign mutations in the genes that were expressed in blood (top 25 percentile) was not significantly different from the ratio in the genes that are not expressed (Supplementary Figure S10; hypergeometric test; P-value > 0.05). Therefore, in somatic tissues the conventional proxies of purifying selection were weak.

We hypothesized that the cell division rate in the stem cell pool might be insufficient for purging deleterious mutations

quickly, leading to the apparent paradox. To test this hypothesis, we applied a published linear birth-immigration-death model (23,24) (see 'Materials and Methods' section for details) to human hematopoiesis to predict how long a HSC clone carrying a somatic mutation could persist in the hematopoietic system under different purifying selection pressure (lack of suitable data limited our ability to extend the analysis to other tissue types). We found that typically, a clone carrying a neutral mutation could persist in the adult HSC pool for 1, 5 and 10 years with probabilities 0.55, 0.25 and 0.18, respectively. Surprisingly, even when a deleterious mutation increased the HSC death rate by 20%, corresponding decrease in probabilities to persist in the pool was modest (i.e. ∼0.54, 0.23 and 0.16, respectively), indicating that such mutations were not purged rapidly from the pool in adults (Figure 4D). While specifics of this model (23,24) can be debated, the key conclusion holds true for broad ranges of parameter regimes. We considered possibilities when mutations affecting replication and differentiation rates were considered in the model (Supplementary Figures S11 and S12); the probability to persist in the pool decreased with an increase in differentiation rate. Key results in Figure 4D were unaffected when we used an alternate model in murine system (36) (Supplementary Text and Supplementary Figure S13). Therefore, it is likely that the deleterious somatic mutations, once acquired, can persist for long time in the stem cell pool, primarily due to slow cell division rate, contributing to the signatures of apparently weak purifying selection. Even though relevant data for other tissues are limited, it appears that the observation can be extended to other tissues as well (Supplementary Text and Supplementary Figure S14). We discuss additional aspects of mutagenesis/selection/differentiation balance in the 'Discussion' section.

DISCUSSION

Taken together, analyzing multiple benign human tissue types, we report the effects of extrinsic and intrinsic mutagenic factors, as well as selection on the landscape of somatic mutations detectable at tissue-level resolution (Figure 4E). To offer a balanced perspective, we note the potential caveats of our analysis. First and importantly, as we did not have access to the biological samples, we could not perform traditional validation using orthogonal methods. But we applied rational filters, and the number of somatic mutations reported here is consistent with that based on theoretical and observed estimates (15,23–26). Second, limited depth of sequencing coverage led us to restrict our analysis to high allelic frequency mutations only. We suspect that an analysis over a broader allele frequency range can provide further critical insights. Third, all the donors were cancer patients, and some of them (e.g. most of those with lung cancer) were smokers. It remains debatable if their non-malignant tissues were truly 'normal'—contributing to the seed and soil hypothesis (37). Therefore, we recommend caution while extending the conclusions to average human populations. Finally, without clinical data we take a conservative approach when discussing functional effects of somatic mutations in apparently benign tissues, and refrain from inferring causality from correlation alone.

Nevertheless, our study provides important insights into mutational signatures in human tissues. For instance, our results suggest that even the mutations acquired later in adulthood can reach high allele frequency in the tissue (e.g. in lung) within the lifetime. This raises a question if life style and environmental exposures not only modify epigenetic makeup, but also lead to discernable changes in genetic makeup of the tissues; this observation, if validated by others, in turn can have implications for late-age diseases. Our cell of origin analysis indicates that many of the mutations detectable at tissue-level resolution were acquired in the long-lived tissue stem or progenitor cells (e.g. in blood). In some cases somatic mutations could directly cause diseased conditions, while in other cases they can modulate disease phenotypes (1). In other cases (e.g. during oncogenesis) the burden of somatic mutations can potentially reduce cellular fitness of healthy cells in benign tissues, increasing selection for adaptive oncogenic events (38). This is probably true also in non-tumor contexts. Even when no disease gene is affected, somatic mutations in both stem cells and their supportive niche cells can reduce overall fitness of these cells, affect cell competition and shape the makeup of the pool of tissue progenitor cells (also tissue phenotype in some cases), during development and aging (29).

The conventional proxies for purifying selection were weak, a trend also observed in cancers. We note that, some of the missense mutations could be recessive in somatic cells, need not always affect protein function or cellular processes in all lineages in the tissue, and that our detection threshold permitted us to only consider high allele frequency somatic mutations. Furthermore, not all such mutations affect cellular fitness and proliferation rates, especially in heterozygous conditions. Nonetheless, the relative abundance of missense mutations, especially those deemed potentially *damaging* and had detectable expression in apparently benign tissues was noteworthy. Our finding is consistent with other reports that (i) there is no evidence for purifying selection on somatic, pathogenic mutations in mitochondria (39), (ii) regulatory mechanisms protect stem cells in benign tissues in embryos from cell competition, and foster cell cooperation (40) (which could allow less fit cells to persist in the population), and (iii) codon usage bias, which also imposes purifying selection, correlates weakly with the tissue-level expression in human (41), and (iv) fluctuating selection (e.g. along developmental hierarchy) can give rise to mutational signatures that could be mistaken as under weak purifying selection or driven by positive selection (42). It is also likely that the effective population size of tissue stem and progenitor cells, their proliferation and differentiation rates, and changing environment in developmental lineages (36,43) might be insufficient for purifying selection to purge these mutations effectively.

Somatic mutations and mosaicism in healthy tissues have implications not only for etiology, diagnosis and treatment of diseases including cancer but also emerging technologies in healthcare. Solid tissues often suffer from field cancerization (44) and may not be 'normal' in molecular characteristics, even when they appear pathologically non-malignant. Low frequency cancer gene somatic mutations have been detected in apparently benign tissues (45,46), which can pose challenges for early detection of malignancies using

liquid biopsy. Furthermore, cell-to-cell genetic variation can be translated into differences in the genetic makeup of the iPS cells derived from the same donor (8), which can bring in unforeseen clinical challenges.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Brent Pedersen, Kyle Smith and Andrii Rohzok for providing helpful comments on the manuscript. S.D. gratefully acknowledges support from the Boettcher Foundation.

FUNDING

Boettcher foundation Webb-Waring Research Grant; NIH (R01CA180175 to J.D.). Funding for open access charge: Boettcher Foundation.
Conflict of interest statement. None declared.

REFERENCES

1. Poduri,A., Evrony,G.D., Cai,X. and Walsh,C.A. (2013) Somatic mutation, genomic variation, and neurological disease. *Science*, **341**, 1237758.
2. Tomasetti,C., Vogelstein,B. and Parmigiani,G. (2013) 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.
3. Aghili,L., Foo,J., De Gregori,J. and De,S. (2014) Patterns of somatically acquired amplifications and deletions in apparently normal tissues of ovarian cancer patients. *Cell Rep.*, **7**, 1310–1319.
4. Jacobs,K.B., Yeager,M., Zhou,W., Wacholder,S., Wang,Z., Rodriguez-Santiago,B., Hutchinson,A., Deng,X., Liu,C., Horner,M.J. *et al.* (2012) Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.*, **44**, 651–658.
5. Jaiswal,S., Fontanillas,P., Flannick,J., Manning,A., Grauman,P.V., Mar,B.G., Lindsley,R.C., Mermel,C.H., Burtt,N., Chavez,A. *et al.* (2014) Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.*, **371**, 2488–2498.
6. Martincorena,I., Roshan,A., Gerstung,M., Ellis,P., Van Loo,P., McLaren,S., Wedge,D.C., Fullam,A., Alexandrov,L.B., Tubio,J.M. *et al.* (2015) Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, **348**, 880–886.
7. O'Huallachain,M., Karczewski,K.J., Weissman,S.M., Urban,A.E. and Snyder,M.P. (2012) Extensive genetic variation in somatic human tissues. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 18018–18023.
8. Abyzov,A., Mariani,J., Palejev,D., Zhang,Y., Haney,M.S., Tomasini,L., Ferrandino,A.F., Rosenberg Belmaker,L.A., Szekely,A., Wilson,M. *et al.* (2012) Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature*, **492**, 438–442.
9. Behjati,S., Huch,M., van Boxtel,R., Karthaus,W., Wedge,D.C., Tamuri,A.U., Martincorena,I., Petljak,M., Alexandrov,L.B., Gundem,G. *et al.* (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, **513**, 422–425.
10. Cai,X., Evrony,G.D., Lehmann,H.S., Elhosary,P.C., Mehta,B.K., Poduri,A. and Walsh,C.A. (2014) Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.*, **8**, 1280–1289.
11. De,S. (2011) Somatic mosaicism in healthy human tissues. *Trends Genet.*, **27**, 217–223.
12. Vijg,J. (2014) Somatic mutations, genome mosaicism, cancer and aging. *Curr. Opin. Genet. Dev.*, **26**, 141–149.
13. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
14. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
15. Holstege,H., Pfeiffer,W., Sie,D., Hulsman,M., Nicholas,T.J., Lee,C.C., Ross,T., Lin,J., Miller,M.A., Ylstra,B. *et al.* (2014) Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res.*, **24**, 733–742.
16. Benaglia,T., Chauveau,D., Hunter,D.R. and Young,D.S. (2009) mixtools: an R package for analyzing mixture models. *J. Stat. Softw.*, **32**, 1–29.
17. Wood,R.D., Mitchell,M. and Lindahl,T. (2005) Human DNA repair genes, 2005. *Mutat. Res.*, **577**, 275–283.
18. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
19. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
20. Sim,N.L., Kumar,P., Hu,J., Henikoff,S., Schneider,G. and Ng,P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452–W457.
21. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6062–6067.
22. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.
23. Catlin,S.N., Guttorp,P. and Abkowitz,J.L. (2005) The kinetics of clonal dominance in myeloproliferative disorders. *Blood*, **106**, 2688–2692.
24. Catlin,S.N., Busque,L., Gale,R.E., Guttorp,P. and Abkowitz,J.L. (2011) The replication rate of human hematopoietic stem cells in vivo. *Blood*, **117**, 4460–4466.
25. Welch,J.S., Ley,T.J., Link,D.C., Miller,C.A., Larson,D.E., Koboldt,D.C., Wartman,L.D., Lamprecht,T.L., Liu,F., Xia,J. *et al.* (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell*, **150**, 264–278.
26. Xie,M., Lu,C., Wang,J., McLellan,M.D., Johnson,K.J., Wendl,M.C., McMichael,J.F., Schmidt,H.K., Yellapantula,V., Miller,C.A. *et al.* (2014) Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.*, **20**, 1472–1478.
27. Cheung,A.M., Nguyen,L.V., Carles,A., Beer,P., Miller,P.H., Knapp,D.J., Dhillon,K., Hirst,M. and Eaves,C.J. (2013) Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood*, **122**, 3129–3137.
28. Verovskaya,E., Broekhuis,M.J., Zwart,E., Ritsema,M., van Os,R., de Haan,G. and Bystrykh,L.V. (2013) Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood*, **122**, 523–532.
29. Stine,R.R. and Matunis,E.L. (2013) Stem cell competition: finding balance in the niche. *Trends Cell Biol.*, **23**, 357–364.
30. Takizawa,H., Regoes,R.R., Boddupalli,C.S., Bonhoeffer,S. and Manz,M.G. (2011) Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. *J. Exp. Med.*, **208**, 273–284.
31. Horvath,S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**, R115.
32. Lee,B.M., Kwack,S.J. and Kim,H.S. (2005) Age-related changes in oxidative DNA damage and benzo(a)pyrene diolepoxide-I (BPDE-I)-DNA adduct levels in human stomach. *J. Toxicol. Environ. Health A*, **68**, 1599–1610.
33. Cancer Genome Atlas Research Network. (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.

34. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Borresen-Dale,A.L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
35. Ostrow,S.L., Barshir,R., DeGregori,J., Yeger-Lotem,E. and Hershberg,R. (2014) Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS Genet.*, **10**, e1004239.
36. Busch,K., Klapproth,K., Barile,M., Flossdorf,M., Holland-Letz,T., Schlenner,S.M., Reth,M., Hofer,T. and Rodewald,H.R. (2015) Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, **518**, 542–546.
37. Fidler,I.J. (2003) The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nature Rev. Cancer*, **3**, 453–458.
38. DeGregori,J. (2013) Challenging the axiom: does the occurrence of oncogenic mutations truly limit cancer development with age? *Oncogene*, **32**, 1869–1875.
39. Greaves,L.C., Elson,J.L., Nooteboom,M., Grady,J.P., Taylor,G.A., Taylor,R.W., Mathers,J.C., Kirkwood,T.B. and Turnbull,D.M. (2012) Comparison of mitochondrial mutation spectra in ageing human colonic epithelium and disease: absence of evidence for purifying selection in somatic mitochondrial DNA point mutations. *PLoS Genet.*, **8**, e1003082.
40. Dejosez,M., Ura,H., Brandt,V.L. and Zwaka,T.P. (2013) Safeguards for cell cooperation in mouse embryogenesis shown by genome-wide cheater screen. *Science*, **341**, 1511–1514.
41. Kanaya,S., Yamada,Y., Kinouchi,M., Kudo,Y. and Ikemura,T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.*, **53**, 290–298.
42. Huerta-Sanchez,E., Durrett,R. and Bustamante,C.D. (2008) Population genetics of polymorphism and divergence under fluctuating selection. *Genetics*, **178**, 325–337.
43. Abkowitz,J.L., Catlin,S.N., McCallie,M.T. and Guttorp,P. (2002) Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. *Blood*, **100**, 2665–2667.
44. Rubin,H. (2011) Fields and field cancerization: the preneoplastic origins of cancer: asymptomatic hyperplastic fields are precursors of neoplasia, and their progression to tumors can be tracked by saturation density in culture. *Bioessays*, **33**, 224–231.
45. Millikan,R., Hulka,B., Thor,A., Zhang,Y., Edgerton,S., Zhang,X., Pei,H., He,M., Wold,L., Melton,L.J. *et al.* (1995) p53 mutations in benign breast tissue. *J. Clin. Oncol.*, **13**, 2293–2300.
46. Bissell,M.J. and Hines,W.C. (2011) Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nat. Med.*, **17**, 320–329.