Title: Conserved patterns of somatic mutations in human blood cells

Authors: L. Alexander Liggett[1,2], Anchal Sharma[3], Subhajyoti De[3], and James DeGregori[1,2,4,5,6*]

***Corresponding Author. Email:** james.degregori@ucdenver.edu

Affiliations:

[1]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA.

[2]Linda Crnic Institute for Down Syndrome, University of Colorado School of Medicine, Aurora, CO 80045, USA.

[3]Rutgers Cancer Institute, New Brunswick, NJ 08901, USA.

[4]Integrated Department of Immunology, University of Colorado School of Medicine, Aurora, CO 80045, USA.

[5]Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA.

[6]Department of Medicine, Section of Hematology, University of Colorado School of Medicine, Aurora, CO 80045, USA.

Abstract

We currently lack an understanding of somatic mutation frequencies and patterns in benign tissues, as studies are often limited to the identification of mutations in clonal expansions (*1*). Using a novel method capable of accurately detecting mutations at single base pair resolution with allele frequencies as rare as 10^{-4}, we find a surprisingly high somatic mutation burden of 50-900 mutations/MB in peripheral blood cells from apparently healthy individuals. Nearly all analyzed sites carry at least one somatic mutation (including known oncogenic mutations) within approximately 20,000 cells. Unexpectedly, mutation patterns and corresponding allele frequencies are highly similar between individuals, age-independent, and lack signatures of selection. We also identified two individuals with patterns of somatic mutation that resemble mismatch repair deficiency, exhibiting mutations that exist at uniformly elevated mutation frequencies. These results demonstrate that somatic mutations, including oncogenic changes, are abundant in healthy human tissue and suggest an unappreciated degree of non-randomness within the processes underlying somatic mutation.

Introduction

The processes involved in somatic mutagenesis are typically regarded as considerably stochastic, and have been incorporated into theories of oncogenesis, aging, and evolution accordingly (*2*, *3*). Nevertheless, it is well known that mutation rates can be influenced by such factors as chromosomal location, nucleotide identity, and sequence context (*4*–*9*). As good example of mutation bias, that represents a substantial number of human point mutations, cytosine deamination within CpG contexts strongly favors C>T point mutations (*10*–*12*). The effect of this bias is furthermore not limited to just the CpG site itself, as mutation rate increases within 10 nucleotides of a CpG dinucleotide (*13*). Moreover, neighboring base pairs can influence the somatic mutability of a nucleotide (*14*, *15*). While many other notable examples of biased mutability have been identified, understanding somatic mutation rates and biases for each nucleotide position within the human genome has been significantly restricted by technological limitations (*9*, *16*–*18*).

Somatic mutations are constantly occurring, yet without clonal expansion, each unique mutation will typically exist at a very low allele frequency (*19*–*22*). This scarcity of somatic mutations makes it challenging to understand how deterministic mutation rates and burdens are, and has provided motivation to improve the sensitivity of existing methods. Technologies such as high-throughput digital droplet PCR (*23*–*25*), COLD-PCR (*26*, *27*), and BEAMing (*28*) have shown promise for rare mutation detection, but are often limited to variant allele frequencies (VAFs) greater than 1 percent or are restricted to assaying only a few mutations at a time. In comparison, sequencing-based approaches can theoretically detect many mutations below a 1 percent allele frequency, but distinguishing true signal from relatively high false positive background has been a significant challenge. These signal to noise difficulties have been somewhat overcome by increasing the depth of sequencing, using clever methods of DNA barcoding, (*19*, *29*) or performing paired strand collapsing (*30*). Despite these advances, sufficiently high false positive rates and low allele capture efficiencies have largely prevented sequencing-based approaches from yielding a comprehensive understanding of mutation rates and biases within the human genome (*19*, *22*, *31*).

A better understanding of somatic mutation rates, could have profound influences on our understanding of somatic evolution and its role in pathogenic processes like oncogenesis (*32*, *33*). The aforementioned technological limitations have made it difficult to study somatic mutation burden and rates in healthy tissue, largely confining measurements of somatic mutation levels to retrospective reconstruction of in vitro (*14*, *34*–*36*) or in vivo (*29*, *36*–*41*) clonally expanded cells. By analyzing clonally expanded cells, these methods are typically confined to analysis of founder cells, and miss further downstream somatic changes. These limitations have left significant gaps in our knowledge of somatic mutation rates within healthy, properly functioning tissue.

Results

To overcome current sequencing limitations, we created FERMI (Fast Extremely Rare Mutation Identification), in which we adapted the amplicon sequencing method of Illumina's TrueSeq Custom Amplicon platform to efficiently capture regions of genomic DNA (gDNA) purified from peripheral blood cells. While targeted sequencing is typically performed on broad regions of DNA, we used DNA probes to target and capture a precise set of 32 genomic regions, each approximately 150bp in length, that span either AML-associated oncogenic mutations or Tier III (non-conserved, non-protein coding and non-repetitive sequence) regions of the human genome.

With a significantly improved probe capture efficiency that yields about 1.2 million unique captures from 1μg of gDNA (see Methods), this approach enabled ultra-deep sequencing of peripheral blood cells. To overcome the false positive signals that often limit the utility of ultra-deep sequencing, we included in our DNA capture probes a 16bp index, containing sequence unique to each probed individual and a 12bp unique molecular identifier (UMI) of randomized DNA unique to each capture (Fig. 1a). Sequencing reads of these capture probes were then sorted by sample index and UMI to produce bins of single cell sequencing which were collapsed to produce largely error-free consensus reads. Captures were only considered if supported by at least 5 sequencing reads, and variants were only included if identified in both paired-end sequences and detected in at least 55% percent of supporting reads for each capture (Fig. 1a and Methods; see also Supplementary Fig. 1).

All probed regions were successfully captured and amplified with some variability in efficiency depending on probe identity (Fig. 1b). To understand assay sensitivity, log-series ratios of one human's gDNA diluted into another's gDNA were analyzed by FERMI. We observed robust quantification of spiked-in single nucleotide polymorphisms (SNPs) with frequencies as rare as 10^{-4} (Fig. 1c). Accurate quantification of SNP frequency can also be made when using strand information to follow dilutions of multiple SNPs located on the same allele (Fig. 1d). For more description of the methods used to maximize the accuracy of FERMI, see *Elimination of false positive signal* in Methods and Supplementary Fig. 1.

To assay somatic mutation burden in peripheral blood and understand how it changes with age, we used FERMI to capture and sequence gDNA from the peripheral blood of 22 apparently healthy donors ranging in age from 0 (cord blood) to 89 years old (Supplementary Table 1). Common and rare germline SNPs could be readily identified by their allele frequencies (Supplementary Fig. S2). In addition, FERMI detected many rare somatic mutations present below 0.3% allele frequencies in these samples. Interestingly, nearly all analyzed sites had at least one somatic mutation across ~20,000 cells in peripheral blood (reflecting ~40,000 captured alleles). These observed mutation rates predict a burden of 50 to 900 mutations per megabase (See

Estimation of mutation burden in Methods), a rate that is much higher than estimations derived from hematopoietic tumor analysis, which typically range from 0.02 - 1mut/Mb (*34*). As leukemias are often of stem or progenitor cell origin (*42*), our elevated estimations suggest that mutation rates might be considerably elevated during production of terminally differentiated cells.

To understand variation within the human population, the variant allele frequency (VAF) for each rare variant was compared between each of the 22 blood donors. Unexpectedly, we found that these rare somatic variants existed at remarkably similar allele frequencies between individuals, across the full sampled age range. These rare VAFs are similar enough between most individuals that inter-individual comparisons for each unique substitution fall along a y=x line (Fig. 2a). We also created an average of the rare VAFs across the 22 donors, and used this for comparison to each individual, which also adhered to a y=x line (R^2 Range = 0.426-0.631, Mean = 0.558) (Fig. 2b-c). Indicative of minimal age-related change in the mechanisms governing leukocyte somatic mutation spectra, the degree of mutation pattern similarity between individuals compared to the population average does not correlate with age (Fig. 2c). These similarities also reproduce in an independent experiment with a separate cohort of blood donors (Supplementary Fig. 3). This lack of age-dependence suggests that most of these mutations were unlikely to have occurred in long-term stem or progenitor cell populations, and instead arose at later stages of hematopoiesis. Furthermore, most variants likely represent multiple independent events rather than clonal expansions, as they are found at similar frequencies on both alleles (Fig. 2d). Consistent with this interpretation, analyses of serial samples from the same donors (Supplementary Fig. 2c) are also highly concordant, suggesting that such mutations probably arise transiently and recurrently in blood cell populations. It thus appears that instead of being semi-random, the aggregate effect of all DNA damage and maintenance processes generates somatic mutations at predictable rates throughout the genome independently of age.

While we observe variants at conserved frequencies across many individuals, previous studies have described age-related clonal expansions of cells containing AML-associated oncogenic changes(*37*–*39*, *41*). Though we observe each queried oncogenic change in every biopsied individual, we do not observe significant age-related increases in the allele frequencies of oncogenic mutations (Fig. 2e and Supplementary Fig. 4). Thus, there was no clear evidence for positive or negative selection accompanying these oncogenic mutations. This inability to observe clonal expansions with age is most likely due to the fact that the average age of the adult individuals within our cohort is only 49 years, with only 5 donors older than 70 years. In addition, oncogenic mutations occurring at later stages of hematopoiesis may be unlikely to result in clonal expansions.

Previous observations suggest disparate mutation rates for each of the four DNA bases(*14*, *37*, *40*, *43*, *44*). Consistent with these observations, we observe nucleotide specific substitution probabilities, with C>T and T>C substitutions being the most common and T>G substitutions being the least common (Fig. 3a). When base change probabilities are analyzed within the context of their two flanking nucleotides (trinucleotide context), significant differences in substitution probability emerge, illustrating a significant impact of nucleotide context on overall mutability (Fig. 3b). While these immediately surrounding bases appear to significantly impact substitution probability, bases further away appear to have relatively little impact (Supplementary Fig. 5a). Also consistent with expectations, CpG positions were more likely to mutate than others in a manner not explained by oversampling of CpG sites (Fig. 3c and Supplementary Fig. 6).

Independent of functional or oncogenic potential, substitutions occur at rates that are uniquely determined by nucleotide position, such that each locus mutates in a highly reproducible manner (Fig 4a-b). Strikingly, a subset of sites shows a highly significant bias to mutate to only one of the possible nucleotides (Supplementary Fig. 7) – across all assayed individuals, these sites mutate to only one of the three possible alternative nucleotides. Even for matching trinucleotide contexts, substitution frequencies can vary, and often fall within just one of two distinct upper and lower VAF clusters (Fig. 5a and Supplementary Table 2). Thus, the substantial variation in mutation probabilities for different positions cannot simply be explained by base bias or trinucleotide context, but likely involves context conferred by other factors like histone and DNA modifications or chromosomal organization. Nonetheless, analyzing the neighboring base contexts for each base change separated into lower-VAF and upper-VAF groups revealed an influence of the flanking bases on mutation frequency for some changes but not others. For example, the immediate flanking bases exerted a much greater influence on the VAFs for C>A changes than C>T changes, and thus explains some but not all variability in mutation frequency (Fig. 5b and Supplementary Fig. 5b).

Possibly indicative of differences in mutational (and possibly selection) processes within cancers, the integrated exome sequencing pan cancer somatic mutation data from The Cancer Genome Atlas (TCGA) exhibits different substitution patterns from those that we find in healthy donor blood (Supplementary Fig. 8). Using the trinucleotide contexts of the substitutions, 7 out of 30 previously identified mutations signatures were identified, and these signatures did not differ significantly across sampled genomic segments (Supplementary Fig. 8).

To explore the ability of FERMI to distinguish perturbations of somatic mutation patterns, gDNA from MMR deficient HCT116 cells (MMR^{MT}) that express truncated and non-functional MLH1 protein was compared to MMR proficient HCT116 cell line gDNA. Providing further validation of our method, across multiple experiments, we observed a substantial increase in VAFs within the MMR^{MT} gDNA when compared to MMR competent HCT116 control (Fig. 6a and Supplementary Fig. 10a). Interestingly, while the mutation spectra of most peripheral blood samples resemble those in other individuals, the spectra from two individuals (samples 2 and 19) possessed a subset of variants that deviated from the population averages, having allele frequencies about twofold higher than average (Fig. 6b-c, and Supplementary Fig. 9). While the magnitude of deviation from mean VAFs was different between the two samples, the identities of the deviating variants were very similar, such that a comparison of VAFs between these two individuals correlate more closely to a y=x line than to the overall population average (Fig. 6d). This consistent deviation in VAFs for these two individuals suggests that the mechanisms governing mutation prevalence can be systematically perturbed in a manner that uniformly alters certain substitution probabilities.

Surprisingly, the substitution VAFs observed within individuals 2 and 19 resembled those altered in the MMR^{MT} HCT116 cells, though the magnitude of these changes were greater in the latter (Fig. 6e). Furthermore, the deviating variants found within individuals 2, 19 and MMR^{MT} samples are not enriched for oncogenic variants (Fig. 6f; shown for individual 2), indicating that deviations are not likely the result of oncogenic selection.

As expected from past studies(*45*), the HCT116 MMR^{MT} gDNA showed an increased prevalence of T>C and T>A substitutions when compared to parental gDNA

(Supplementary Fig. 10). Peripheral blood gDNA from individuals 2 and 19 also exhibited similarly increased rates of T>C and T>A substitutions (Fig. 6g-h and Supplementary Fig. 9). Thus, these two individuals appear to exhibit a mild MMR deficiency. In support of the results, individuals 2 and 19 show the same increased rates of substitutions across two experiments, with strong reproducibility in mutation patterns (Supplementary Fig. 9j). The systematic and reproducible variance from the typical mutational pattern for these two individuals and the MMR^{MT} HCT116 cells also serves as validation of the specificity of FERMI to accurately detect variants and their frequencies. More importantly, the identification of two individuals with altered somatic mutation patterns out of only 35 individuals may indicate that systematic somatic mutation deviations from typical mutational profiles may be relatively common in the human population.

Discussion

In this study, we created a unique method of measuring levels of ultra rare somatic mutations and mutational burden within human blood. Use of this method gave rise to several key findings. First, we found an unexpectedly high somatic mutation burden within putatively healthy individuals, where cells contained 100-1000 times more mutations than previous measurements in stem cells and even most cancers. As tumors may often originate from stem cells, retrospective analysis of high frequency variants may largely reflect stem cell mutation burden. Together with measurements derived from healthy stem cells, our higher observed mutation burdens in mature blood cells could reflect unique use of mutation reducing processes within stem cells such as lower cell division rates, reduced exposure to oxidative stress, higher efflux pump activity, and perhaps better DNA repair. The second important finding was that all probed oncogenic changes were observed in each evaluated individual without evidence of either positive or negative selection, suggesting that oncogenic mutations occurring in short-lived, lineage-committed cells pose minimal risk. Moreover, these results indicate that oncogenic mutations are not uniformly under positive selection in normal tissues.

The third important observation was the surprising degree of similarity between the somatic mutation patterns of different individuals. If somatic mutations occurred in a largely stochastic manner, it may be logical to expect striking differences between individual somatic mutation patterns. Yet within our cohort only two individuals exhibited noticeable differences from the others. Furthermore, while nucleotide context is known to influence the mutability of genomic loci, we find that each nucleotide locus carries with it a uniquely determined mutability rate for each possible substitution. These mutability rates are conserved across nearly all measured individuals, and appear responsible for the observed similarities in somatic mutation. This would suggest that while somatic mutagenesis is often seen as a largely random process, in reality, it appears to be governed by a number of complex and highly deterministic factors.

Human mutation rates have long been an area of study, but technological limitations have largely necessitated that they be indirectly measured through clonal expansions of isolated healthy cells, in tumor cells, or from germline mutation rates across generations (*35*–*37*). While DNA sequencing based methods allow for the observation of ultra rare mutations, if enough DNA is sequenced to reach the allele frequencies present in somatic cells, false positive rates tend to climb high enough to obscure true rare mutations. We solved these problems with a barcoding system that allowed each captured genomic allele to be distinguished from that of other captures, providing near single cell sequencing resolution to bulk sequencing experiments. Furthermore, we sufficiently improved DNA capture efficiencies to allow capture of millions of unique alleles from each blood biopsy. This high allelic capture rate enabled reliable detection of mutations at rare enough allele frequencies that spontaneous somatic mutations could be observed. This sequencing strategy revealed somatic mutation loads per cell that are orders of magnitude higher than those measured in hematopoietic stem or progenitor cells (following clonal expansion) (*34*). Given our estimates of 50 to 900 mutations/MB in the average mature leukocyte, this burden would suggest a mutation rate between 10^{-6} to 10^{-5} mutations per nucleotide per division, respectively, if one assumes that a mature leukocyte is the product of approximately 100 cell divisions. Not only is this mutation rate substantially higher than those measured in progenitor cells, even in mature skin cells, mutation rates are only about 6 mutations/MB (*40*). The disparity between mutation rates for cell divisions in short-term progenitors (leading to terminal cells) and cell divisions in stem and germ cells may reflect the importance of investing more heavily in genome damage avoidance and repair within stem and germ cell populations. Furthermore, the accuracy of FERMI may facilitate a better understanding of the extent of somatic evolution within tumors. As previously elaborated, typical studies are confined to retrospective study of early driver mutations within clonogenic expansions, but are technologically limited from understanding later mutation acquisition within tumors (subsequent to the most recent bottlenecks). FERMI could be used to better understand how cancers evolve, particularly if leveraged during periodic sampling of malignancies.

Our studies demonstrated that within about 20,000 blood cells (2-5 μl of blood) all queried oncogenic mutations were present in each biopsied individual. While previous studies have demonstrated clonal expansions of some oncogenically initiated cells in a fraction of elderly individuals (*37*–*39*), we observe no such effect. While this is likely due to insufficient old-age samples, we were surprised to find that oncogenic mutations are ubiquitous in even very young individuals and at conserved frequency regardless of age. This is consistent with the frequent detection of oncogenic mutations in individuals over 50 in a previous study (*29*). This finding is bolstered by, and may help explain the previously reported commonality of oncogenically-initiated clonal expansions in sun exposed skin (*46*). As we are largely sampling terminally differentiated cells, we conclude that oncogenic mutations are being reliably generated during the production of mature cells throughout human life at consistent rates.

Given the common presence of oncogenic mutations in normal tissues, numerous hurdles clearly exist that prevent further cancer evolution, including intrinsic tumor suppressor pathways such as senescence and the hierarchical organization of tissues (*47*, *48*). In cancers like acute myeloid, chronic lymphoid and chronic myeloid leukemias, which have been shown to initiate in hematopoietic stem cells (*49*–*52*), the small numbers and low division rates of these target stem cells should serve as a barrier to oncogenesis. Our results also highlight the importance of tissue maintenance mechanisms, which can maintain functionality despite mutation accumulation, in limiting and delaying both cancer and aging (*47*, *53*). Finally, the prevalence of oncogenic mutations in benign tissues may introduce important challenges to early detection and monitoring of cancer progression.

Additionally, our results indicate that cells carrying oncogenic and other novel epitope-generating mutations are not readily eliminated by the immune system, as might be expected. This is perhaps due to insufficient accompaniment by damage signals like cytoplasmic DNA, or interferon and interleukin signaling (*54*). Furthermore, it is possible that this frequent generation of non-synonymous mutations during human life acts as a tolerizing mechanism that may limit the effectiveness of the immune system in attacking and eliminating tumors or oncogenic expansions.

From previous studies, we expected to observe some bias in mutational frequencies based on sequence context, but that the overall somatic mutation profile would be highly random and unique to an individual at a particular moment in time. Instead, what we found was an incredible degree of similarity between the somatic mutation profiles of each biopsied individual. We show that somatic mutation burden is so highly conserved that each observed substitution exists at very similar frequencies within most biopsied individuals. Furthermore, the manner in which a nucleotide mutates appears to be highly dependent on its particular base location. We expect this dependency reflects the impact of surrounding nucleotides, chromosome context, and epigenetic profile. From these observations (extrapolated genome-wide), we hypothesize that nearly all somatic mutation is predictable and deterministic.

Finally, we observe two individuals whose somatic mutation burden deviates from the others. Surprisingly, both appeared to closely resemble the patterns created by mismatch repair deficiency. With only two samples displaying such a phenotype, it is challenging to understand its populational prevalence, but these results suggest that deviation from typical mutation frequencies may be relatively common. While we already know of some differences in human mutation patterns (*55–57*), if mutation incidence rates can be significantly increased or decreased without affecting cancer or aging rates would indicate that the human body's tolerance for mutations may be greater than previously appreciated.

1. I. Martincorena, P. J. Campbell, Somatic mutation in cancer and normal cells. *Science*. **349**, 1483–1489 (2015).

2. J. H. Bielas, L. A. Loeb, Quantification of random genomic mutations. *Nat. Methods*. **2**, 285–290 (2005).

3. C. Tomasetti, L. Li, B. Vogelstein, Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*. **355**, 1330–1334 (2017).

4. S. Benzer, ON THE TOPOGRAPHY OF THE GENETIC FINE STRUCTURE. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 403–415 (1961).

5. D. J. Gaffney, P. D. Keightley, The scale of mutational variation in the murid genome. *Genome Res.* **15**, 1086–1094 (2005).

6. M. J. Lercher, E. J. B. Williams, L. D. Hurst, Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**, 2032–2039 (2001).

7. M. W. Nachman, S. L. Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics*. **156**, 297–304 (2000).

8. D. G. Hwang, P. Green, Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13994–14001 (2004).

9. A. Hodgkinson, E. Ladoukakis, A. Eyre-Walker, Cryptic variation in the human mutation rate. *PLoS Biol.* **7**, e1000027 (2009).

10. K. J. Fryxell, W.-J. Moon, CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* **22**, 650–658 (2005).

11. L. A. Frederico, T. A. Kunkel, B. R. Shaw, A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*. **29**, 2532–2537 (1990).

12. T. Lindahl, B. Nyberg, Rate of depurination of native deoxyribonucleic acid. *Biochemistry*. **11**, 3610–3618 (1972).

13. W. Qu *et al.*, Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns. *Genome Res.* **22**, 1419–1425 (2012).

14. F. Blokzijl *et al.*, Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*. **538**, 260–264 (2016).

15. Y. S. Ju *et al.*, Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*. **543**, 714–718 (2017).

16. P. L. F. Johnson, I. Hellmann, Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol. Evol.* **3**, 842–850 (2011).

17. V. B. Seplyarskiy, P. Kharchenko, A. S. Kondrashov, G. A. Bazykin, Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Mol. Biol. Evol.* **29**, 1943–1955 (2012).

18. A. Y. Panchin, S. I. Mitrofanov, A. V. Alexeevski, S. A. Spirin, Y. V. Panchin, New words in human mutagenesis. *BMC Bioinformatics*. **12**, 268 (2011).

19. J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O'Roak, J. Shendure, Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854 (2013).

20. J. L. Preston *et al.*, High-specificity detection of rare alleles with Paired-End Low Error Sequencing (PELE-Seq). *BMC Genomics*. **17**, 464 (2016).

21. T.-H. Zhang, N. C. Wu, R. Sun, A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC Genomics*, 1–9 (2016).

22. M. W. Schmitt *et al.*, Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods*, 1–4 (2015).

23. B. J. Hindson *et al.*, High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83**, 8604–8610 (2011).

24. P. J. Sykes *et al.*, Quantitation of targets for PCR by use of limiting dilution. *Biotechniques*. **13**, 444–449 (1992).

25. B. Vogelstein, K. W. Kinzler, Digital PCR. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9236–9241 (1999).

26. C. A. Milbury, M. Correll, J. Quackenbush, R. Rubio, G. M. Makrigiorgos, COLD-PCR enrichment of rare cancer mutations prior to targeted amplicon resequencing. *Clin. Chem.* **58**, 580–589 (2012).

27. J. Li *et al.*, Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nat. Med.* **14**, 579–584 (2008).

28. D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, B. Vogelstein, Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8817–8822 (2003).

29. A. L. Young, G. A. Challen, B. M. Birmann, T. E. Druley, Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).

30. S. R. Kennedy *et al.*, Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).

31. L. Chen, P. Liu, T. C. Evans Jr, L. M. Ettwiller, DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*. **355**, 752–756 (2017).

32. A. Stoltzfus, D. M. McCandlish, Mutational Biases Influence Parallel Adaptation. *Mol. Biol. Evol.* **34**, 2163–2172 (2017).

33. V. L. Cannataro, S. G. Gaffney, J. P. Townsend, Effect sizes of somatic mutations in cancer. *bioRxiv* (2018), p. 229724.

34. J. S. Welch *et al.*, The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell*. **150**, 264–278 (2012).

35. J. Vijg, X. Dong, L. Zhang, A high-fidelity method for genomic sequencing of single somatic cells reveals a very high mutational burden. *Exp. Biol. Med.* . **242**, 1318–1324 (2017).

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).

37. S. Jaiswal *et al.*, Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.*, 1–11 (2014).

38. G. Genovese *et al.*, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).

39. M. Xie *et al.*, Age-related mutations associated with clonal hematopoietic

expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).

40. I. Martincorena *et al.*, Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*. **348**, 880–886 (2015).

41. T. McKerrell *et al.*, Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).

42. M. R. Corces-Zimmerman, R. Majeti, Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia*. **28**, 2276–2282 (2014).

43. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer. *Nature*. **500**, 415–421 (2013).

44. L. B. Alexandrov *et al.*, Mutational signatures associated with tobacco smoking in human cancer. *Science*. **354**, 618–622 (2016).

45. H. Zhao *et al.*, Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. *eLife Sciences*. **3**, e02725 (2014).

46. I. Martincorena *et al.*, High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*. **348**, 880–886 (2015).

47. J. DeGregori, Evolved tumor suppression: why are we so good at not getting cancer? *Cancer Res.* **71**, 3739–3744 (2011).

48. J. DeGregori, Challenging the axiom: does the occurrence of oncogenic mutations truly limit cancer development with age? *Oncogene*. **32**, 1869–1875 (2013).

49. P. J. Fialkow, S. M. Gartler, A. Yoshida, Clonal origin of chronic myelocytic leukemia in man. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 1468–1471 (1967).

50. M. Jan *et al.*, Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4**, 149ra118–149ra118 (2012).

51. Y. Kikushige *et al.*, Self-renewing hematopoietic stem cell is the primary target in pathogenesis of human chronic lymphocytic leukemia. *Cancer Cell*. **20**, 246–259 (2011).

52. T. Miyamoto, I. L. Weissman, K. Akashi, AML1/ETO-expressing nonleukemic stem

cells in acute myelogenous leukemia with 8;21 chromosomal translocation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7521–7526 (2000).

53. A. Rozhok, J. DeGregori, Somatic maintenance alters selection acting on mutation rate. *bioRxiv* (2018), p. 181065.

54. G. N. Barber, STING: infection, inflammation and cancer. *Nat. Rev. Immunol.* **15**, 760–770 (2015).

55. K. Harris, Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3439–3444 (2015).

56. K. Harris, J. K. Pritchard, Rapid evolution of the human mutation spectrum. *Elife*. **6** (2017), doi:10.7554/eLife.24284.

57. I. Mathieson, D. Reich, Differences in the rare variant spectrum among human populations. *PLoS Genet.* **13**, e1006581 (2017).

58. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

59. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

60. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921 (2001).

61. P. A. Fujita *et al.*, The UCSC genome browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2010).

62. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012), (available at http://arxiv.org/abs/1207.3907).

63. A. Tan, G. R. Abecasis, H. M. Kang, Unified representation of genetic variants. *Bioinformatics*. **31**, 2202–2204 (2015).

64. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, C. Swanton, DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).

65. J. S. Gehring, B. Fischer, M. Lawrence, W. Huber, SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*. **31**,

3673–3675 (2015).

66. F. Blokzijl, R. Janssen, R. Van Boxtel, E. Cuppen, MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues. *bioRxiv* (2016), p. 071761.

Acknowledgments: We would like to thank Ruth Hershberg of Technion University and Jay Hesselberth and Robert Sclafani of the University of Colorado School of Medicine for useful suggestions and for review of the manuscript. These studies were supported by grants from the National Cancer Institute (R01CA180175 to J.D.), NIH/NCATS Colorado CTSI Grant Number UL1TR001082CU (seed grant to J.D.), F31CA196231 (to L.A.L.), the Linda Crnic Institute for Down Syndrome (to J.D. and L.A.L.), and P30-CA072720 (to A.S. and S.D.). The research utilized services of the Cancer Center Genomics Shared Resource, which is supported in part by NIH grant P30-CA46934. L.A.L. and J.D. developed the concept of this project, planned the experiments, analyzed results, and wrote the manuscript. L.A.L. processed and prepared samples from blood biopsy to sequencing, and wrote the bioinformatics software used for analysis. A.S. and S.D. analyzed results, and contributed to writing of the manuscript.

Fig. 1

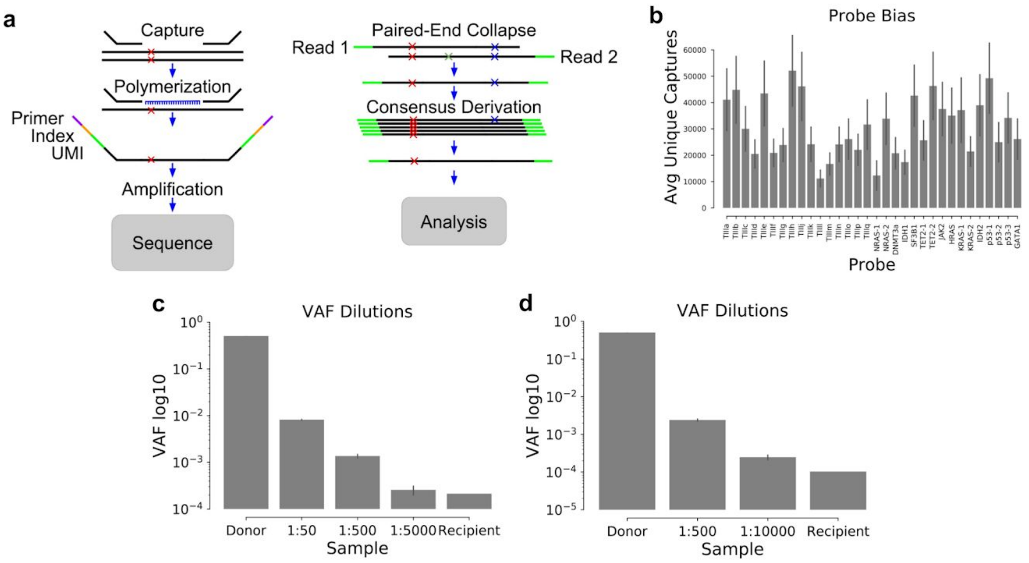


Fig. 1| Amplicon sequencing accurately detects mutation allele frequencies as rare as 1/10,000. a, Graphical depiction of gDNA capture and analysis method. **b**, Capture efficiencies vary in a probe dependent manner. **c**, Accurate detection of a single heterozygous SNP in gDNA from one individual diluted into gDNA from another (without this germline SNP). Dilutions were performed to bring final variant allele frequency to as low as 1/10,000. **d**, Accurate detection of three linked SNPs found within the same allele diluted as in c. For c and d, error shown is standard deviation.

Fig. 2

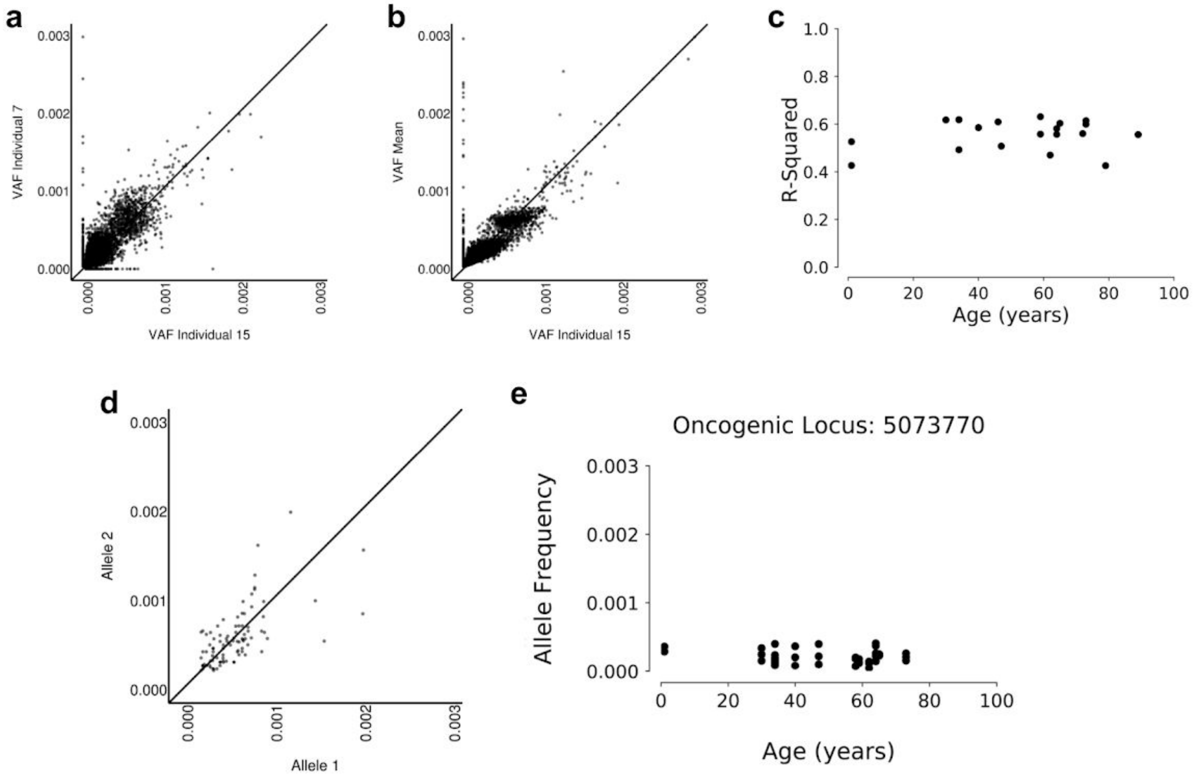


Fig. 2| Mutations exist at conserved frequencies independently of age. a, Comparison of VAFs of identified variants within a 34 year old (x-axis) and 62 year old (y-axis); R^2 = 0.408211, p=0.000. Unless otherwise noted, R^2 values are calculated for all VAFs less than 0.003, which includes almost all somatic variants, but excludes germline variants. **b**, VAFs from a 34 year old (x-axis) compared to mean VAFs calculated from individuals ranging in ages from newborn to 89 years of age (n=22); R^2 = 0.590412, p=0.000. **c**, R^2 values for each individual, compared to the 22-sample VAF mean, plotted by age of the individuals. **d**, Respective frequencies of matched variants on opposite alleles, as determined by linkage to germline SNPs. **e**, Oncogenic VAFs of JAK2 c.1849G>T p.V617F chr9:5073770 plotted as a function of donor age does not reveal evidence of clonal expansions.

Fig. 3

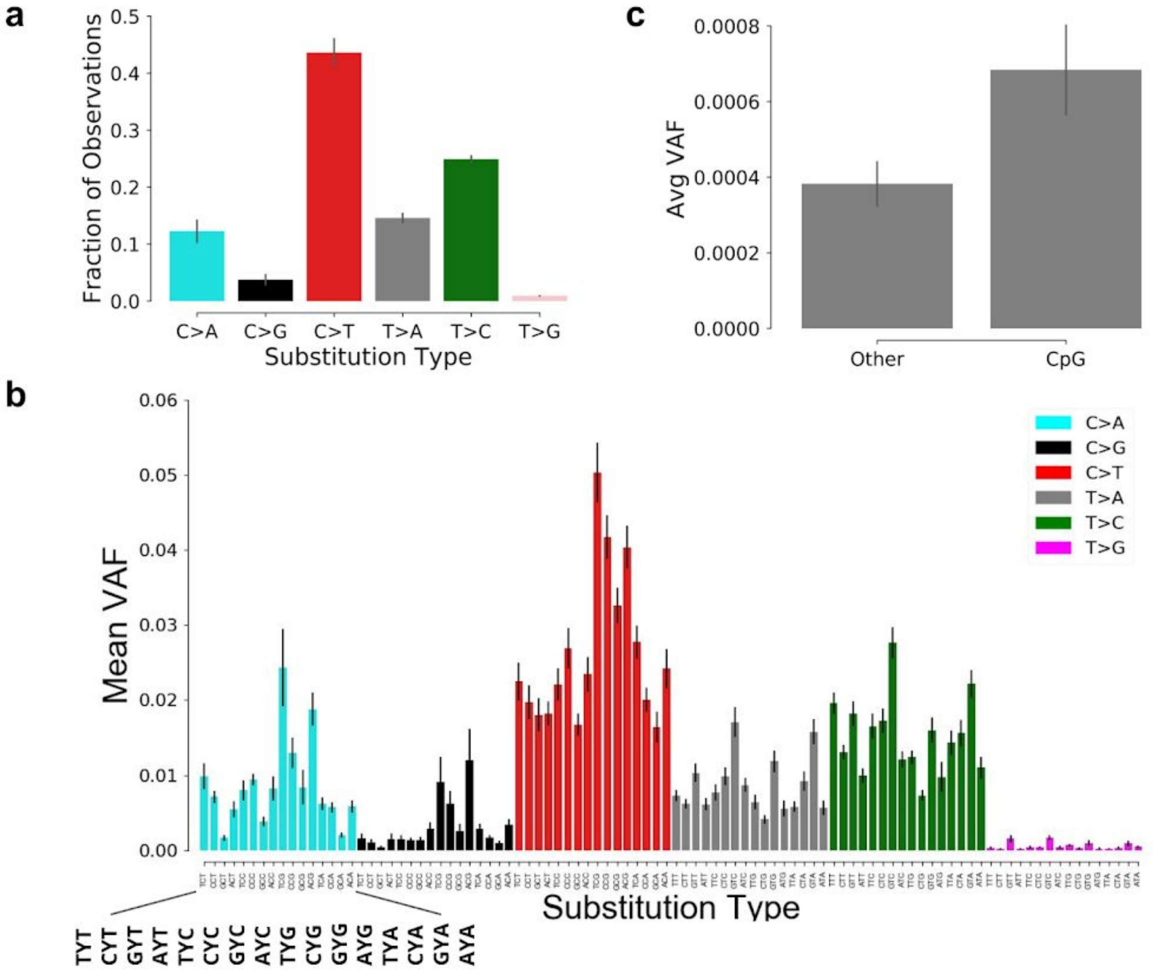


Fig. 3| Sequence context impacts nucleotide mutability. a, Relative rates of each substitution type. Substitutions are quantified by number of supporting unique captures and normalized to 1 as a fraction of all six substitution types (error is standard deviation across individuals). **b**, Relative rates of each substitution type classified by trinucleotide context. Substitutions are quantified by number of supporting unique captures normalized to 1 as a fraction of all other substitution types (error represents standard deviation across individuals). **c**, Separation of C>T changes into CpG and non-CpG sites, showing the average number of variants by capture for each position.

Fig. 4

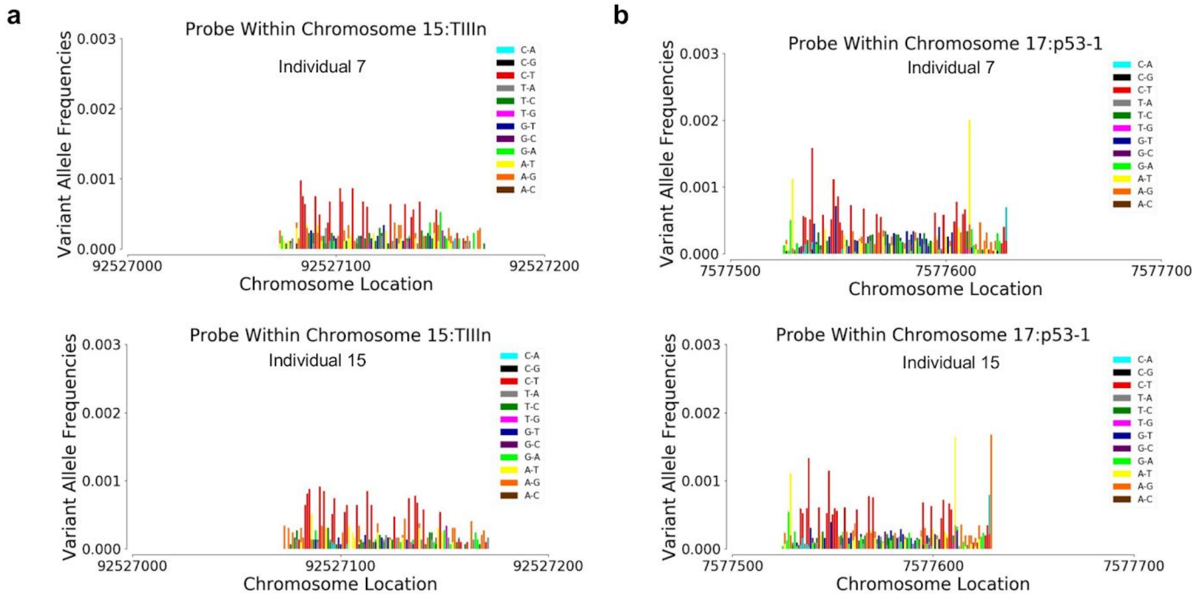


Fig. 4| Loci mutate with specific patterns. a-b, All identified base substitutions within two different probed regions are plotted by their position and allele frequencies for individuals 7 and 15 (representative of all other individuals, with greater deviation observed for individuals 2 and 19 as described below), revealing highly reproducible patterns.

Fig. 5

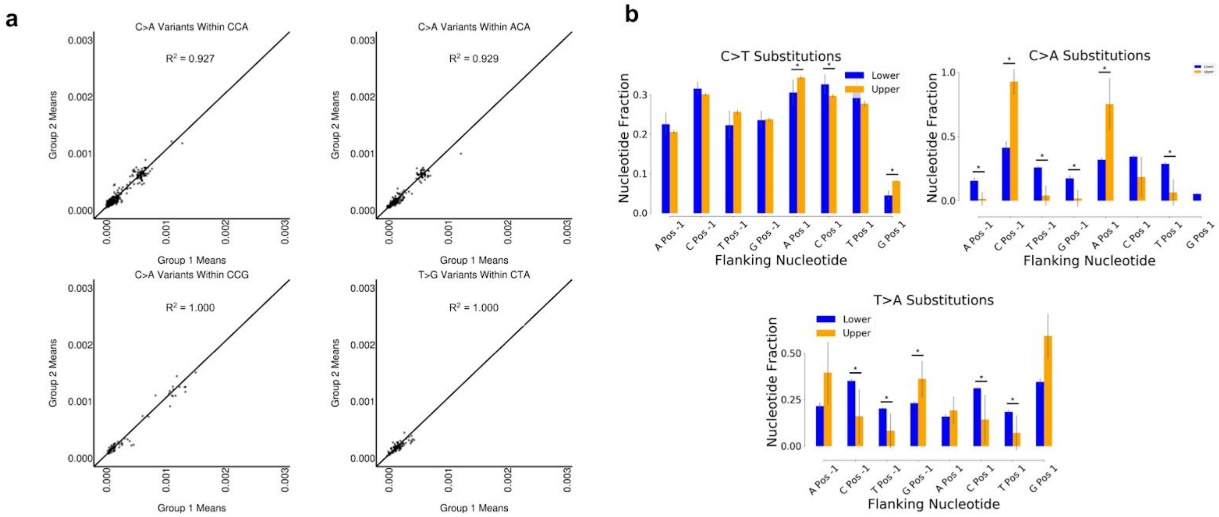


Fig. 5| Trinucleotide context variably impacts variant frequency. a, All individuals were split into two similar groups with similar age distributions. Variants sorted by nucleotide change and triplet context were plotted by VAF. **b**, In order to better understand why some bases were mutated at high frequency, and others were mutated at low frequency, as observed in Fig 5a, the observation frequencies of the bases immediately upstream and downstream were compared for the same change found at high VAFs (upper) and at low VAFs (lower). Using standard deviation between individuals (n=22) and Bonferroni correction for multiple comparison, there appear to be a number of significant differences in the makeup of the upstream and downstream bases for particular base changes. Shown here are only three of the possible six substitutions, as the other three are supported by too few variants in the high VAF population to result in any significant differences between the upper and lower populations.

Fig. 6

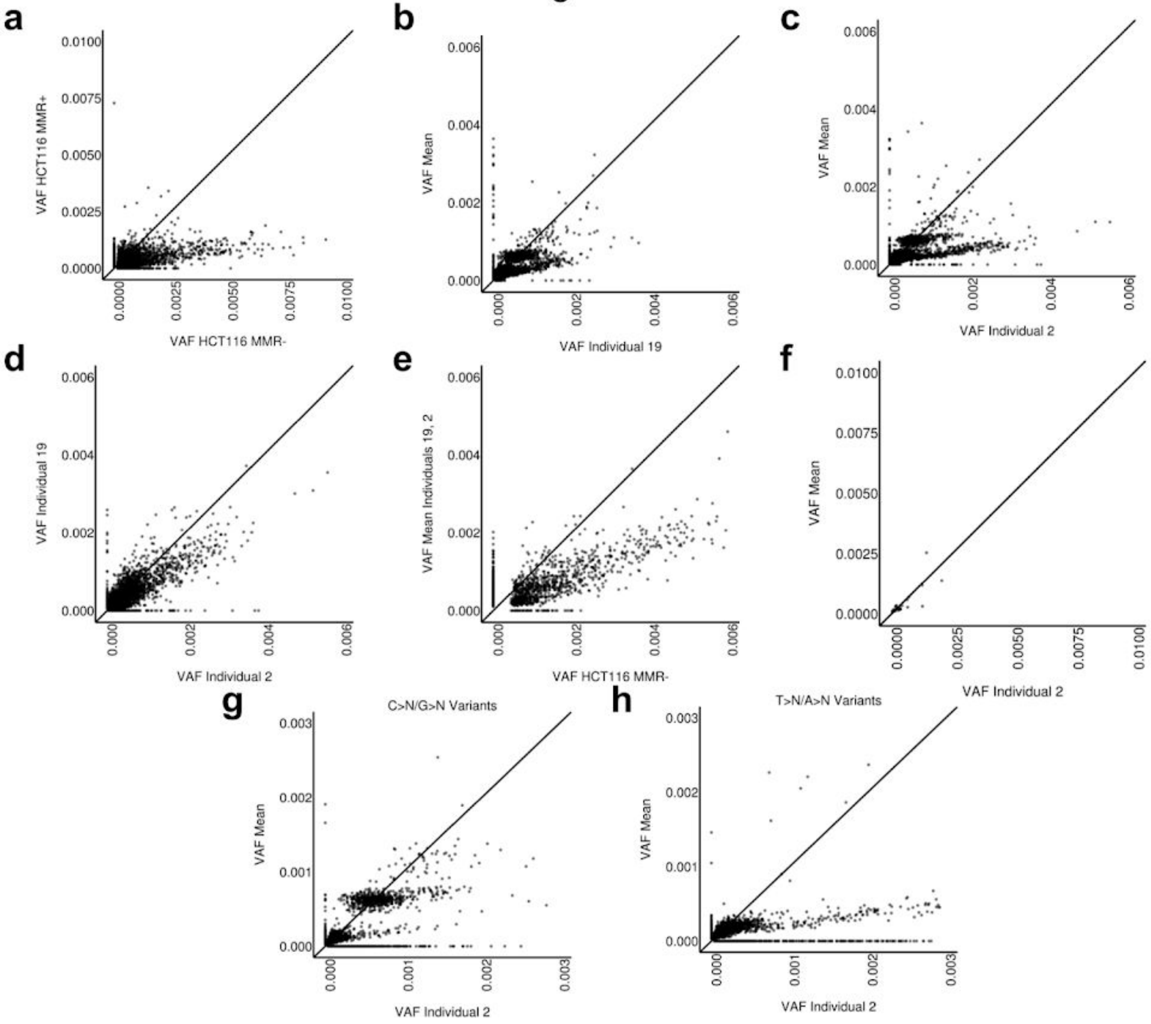


Fig. 6| Individuals can systematically deviate from the population average. a, Comparing VAFs in HCT116 MMR$^+$ vs MMR^{MT} cells reveals an increase in frequencies for many of the observed variants in MMR^{MT} cells (R^2 = 0.211479). **b**, Blood from a 73 year old person (individual #19) compared to the mean VAFs reveals a deviating population of variants that exist at an increased frequency compared with average VAFs (R^2 = 0.387125). **c**, A cord blood sample (individual #2) also shows a subset of variants with higher frequencies than in the average (R^2 = 0.278250). **d**, Comparison of VAFs from individual #2 vs individual #19 reveals that the deviating variants are at the same positions, causing the comparison to fall close to the y=x line (R^2 = 0.613542). **e**, Plotting the mean for VAFs from individuals #2 and #19 versus VAFs from MMR^{MT} HCT116 cells reveals that the variants within the blood are the same as those found within the MMR^{MT} cell line. While variant frequencies are higher in the MMR^{MT} cell line, the proportional change for different deviating variants are similar (R^2 = 0.587474). **f**, Variants detected in individual #2 are not enriched for oncogenic changes (plotted as in Fig. 6c, but only for oncogenic changes). **g**, Plot of only C>N/G>N variants shows relative similarity between individual #2 and the average for all other individuals (R^2 = 0.350623). **i**, Plot of only T>N/A>N variants reveals that the majority of deviating variants for individual #2 are substitutions affecting T or A (R-Squared = 0.040712).

Supplementary Materials:

Materials and Methods

Supplementary Text

Supplementary Figures 1-10

Table S1

References (36-44)

Materials and Methods:

Amplicon Design

Amplicon probes for targeted annealing regions were created using the Illumina Custom Amplicon DesignStudio (https://designstudio.illumina.com/). UMIs were then added to the designed probe regions and generated by IDT using machine mixing for the randomized DNA. Probes were PAGE purified by IDT. All probes are listed below along with binding locations and expected lengths of captured sequence.

Gene	Probe Up	Probe Down	Probe Start	Probe End	Length
JAK2	AGTTTACACTGACACCTAGCTGTGATC	CCATAATTTAAAACCAAATGCTTGTGAGAA	chr9:5073733	chr9:5073887	155
TP53-1	TCATCTTGGGCCTGTGTTATCTCCTA	ATCCTCACCATCATCACACTGGAAGAC	chr17:7577504	chr17:7577635	132
TP53-2	CCCTCAACAAGATGTTTTGCCAACTG	ATGAGCGCTGCTCAGATAGCGATGGT	chr17:7578369	chr17:7578544	176

TP53-3	GGACAGGTAGGACCTGATTTCCTTACT	TGTCCTGGGAGAGACCGGCGCACAGA	chr17:7577084	chr17:7577214	131
NRAS-1	CAATAGCATTGCATTCCCTGTGGTTTT	GTACAGTGCCATGAGAGACCAATACAT	chr1:115256496	chr1:115256680	185
NRAS-2	GAAGGTCACACTAGGGTTTTCATTTCC	AAAAGCGCACTGACAATCCAGCTA	chr1:115258713	chr1:115258897	185
HRAS	TCCTTGGCAGGTGGGGCAGGAGACCC	GCAAGAGTGCGCTGACCATCCA	chr11:534258	chr1:534385	128
KRAS-1	AGGTACTGGTGGAGTATTTGATAGTGT	CAAGAGTGCCTTGACGATACAGCTAATT	chr12:25398247	chr12:25398415	169
KRAS-2	GACTGTGTTTCTCCCTTCTCAGGATTC	TACAGTGCAATGAGGGACCAGTACATG	chr12:25380242	chr12:25380368	127
TET2-1	CCATGTTTTGGCTCATTCATGCTCTTA	ACGGCCACTCCCCCAATGTCAG	chr4:106197237	chr4:106197405	169
TET2-2	CTTTTGAAAGAGTGCCACTTGGTGTCT	GGTGATGGTATCAGGAATGGACTTAGTC	chr4:106155137	chr4:106155275	139
DNMT3A	TGTGTGGTTAGACGGCTTCCGGGCA	AGGCAGAGACTGCTGGGCCGGTCA	chr2:25457211	chr2:25457364	154
IDH1	CAAATGTGGAAATCACCAAATGGCACC	TGGGGATCAAGTAAGTCATGTTGGCA	chr2:209113077	chr2:209113239	163
IDH2	GAAGAAGATGTGGAAAAGTCCCAATGG	CATGGCGACCAGGTAGGCCAGG	chr15:90631809	chr15:90631969	161
GATA1	CTTCCAGCCATTTCTGAGATATCCTCA	CAGCTGCAGCGGTGGCTGTGCT	chrX:48649667	chrX:48649849	183
SF3B1	GTGAACATATTCTGCAGTTTGGCTGAA	ACCATCAGTGCTTTGGCCATTGC	chr2:198266803	chr2:198266967	165
TIIIA	CATCTATTCTGTGCTAGGCATTGTGTG	CAGACCTAGCATCTGTGCCAGAC	chr1:115227814	chr1:115227978	165
TIIIB	CAGTCTGGGTTTTGGAGCAATGATATC	GCAGTGAGCTCAGCCTTGATTTT	chr2:223190674	chr2:223190820	147
TIIIC	CCTGGTGCTTAGTC	AGTCTTCTATAATGC	chr2:229041101	chr2:229041289	189

	CTGTTCTGAAATT	CACAACCTGTAT			
TIIID	GAACAGAACACTTGGTAGTTGACCATG	AGACAGGGAACTGGCATGAAGAGTTT	chr4:110541172	chr4:110541302	131
TIIIE	GCCTAGAACAGGCACCATACATTCAAT	AGATGGTGTTGCTGTGCCGGATAGGAG	chr4:112997214	chr4:112997386	173
TIIIF	TGGCACTATGTGGAGATGTTAGTACAG	GGATGTTGGTGCTATCAGTAGCCATA	chr4:121167756	chr4:121167884	129
TIIIG	CTCTAGGCTTAGTGGTCAAGGAATGAA	AGAAGCAGGACTGTGCTTCCAAACAA	chr4:123547743	chr4:123547901	159
TIIIH	CTTGGTGGTAGCCTAGGCAGTAATTAA	CACGTGGTTGGGAAGAGAAAGTG	chr4:124428637	chr4:124428767	131
TIIIJ	TTCTATAGCACTGGTGACCAGGACACT	CTGGCCACAGTGCCTGGTTTCC	chr11:2126256	chr11:2126420	165
TIIIK	AGACAGGAGGAAGGAGCAATTCAGAAG	CATGGAGATCTCGTCCCCTCAGA	chr11:2389983	chr11:2390171	189
TIIIL	TAGGCCAGAAAACACACAGTGTCGGG	AACTCCGGTAAGTGGCGGGTGGGGGT	chr11:2593889	chr11:2594074	186
TIIIM	ATCTGGGAACAGACCTTCCTCAGGCAT	GTTCTAAGTTACTCTGTGTACTTGACT	chr11:11486596	chr11:11486728	133
TIIIN	AGCCTAGTTACCATAGACGGATTCAAC	GAATATCTTCTAACTGGACTTAGAAAACC	chr15:92527052	chr15:92527176	125
TIIIO	CCAACATGTTCTAAATTCTGGCCACAG	TGGGTCTCAGCCATCCCATTACTG	chr16:73379656	chr16:73379832	177
TIIIP	CTAACATCTCACTTCTACCCTACGCTA	TAAGTGCCCACTACCCCATCCTTAAT	chr16:82455026	chr16:82455164	139
TIIIQ	TCATGACCCAGGCCTCCCAGAACTGAG	ATCTGTGAAGCCGGAGTGAAAACAAC	chr16:85949137	chr16:85949299	163

Genomic DNA Isolation

Human blood samples were purchased from the Bonfils Blood Center Headquarters of Denver Colorado. Our use of these samples was determined to be “Not Human Subjects” by our Institutional Review Board. Biopsies were collected as unfractionated whole blood from apparently healthy donors, though samples were not tested for infection. Samples were approximately 10 mL in volume, and collected in BD Vacutainer spray-coated EDTA tubes. Following collection, samples were stored at 4°C until processing, which occurred within 5 hours of donation. To remove plasma from the blood, samples were put in 50 mL conical tubes (Corning #430828) and centrifuged for 10 minutes at 515 rcf. Following centrifugation, plasma was aspirated and 200 mL of 4°C hemolytic buffer (8.3g NH_4Cl, 1.0g $NaHCO_3$, 0.04 Na_2 in 1L ddH_2O) was added to the samples and incubated at 4°C for 10 minutes. Hemolyzed cells were centrifuged at 515 rcf for 10 minutes, supernatant was aspirated, and pellet was washed with 200 mL of 4°C PBS. Washed cells were centrifuged for at 515rcf for 10 minutes, from which gDNA was extracted using a DNeasy Blood & Tissue Kit (Qiagen REF 69504).

Amplicon Capture

For amplicon capture from gDNA, we modified the Illumina protocol called “Preparing Libraries for Sequencing on the MiSeq” (Illumina Part #15039740 Revision D). DNA was quantified with a NanoDrop 2000c (ThermoFisher Catalog #ND-2000C). 500ng of input DNA in 15µl was used for each reaction instead of the recommended quantities. In place of 5µl of Illumina 'CAT' amplicons, 5µl of 4500ng/µl of our amplicons were used. During the hybridization reaction, after gDNA and amplicon reaction mixture was prepared, sealed, and centrifuged as instructed, gDNA was melted for 10 minutes at 95°C in a heat block (SciGene Hybex Microsample Incubator Catalog #1057-30-O). Heat block temperature was then set to 60°C, allowed to passively cool from 95°C and incubated for 24hr. Following incubation, the heat block was set to 40°C and allowed to passively cool for 1hr. The extension-ligation reaction was prepared using 90 µl of ELM4 master mix per sample and incubated at 37°C for 24hr. PCR amplification was performed at recommended temperatures and times for 29 cycles. Successful amplification was confirmed immediately following PCR amplification using a Bioanalyzer (Agilent Genomics 2200 Tapestation Catalog #G2964-90002, High Sensitivity D1000 ScreenTape Catalog #5067-5584, High Sensitivity D1000 Reagents Catalog #5067-5585). PCR cleanup was then performed as described in Illumina's protocol using 45 µl of AMPure XP beads. Libraries were then normalized for sequencing using the Illumina KapaBiosystems qPCR kit (KapaBiosystems Reference # 07960336001).

Sequencing

Prepared libraries were pooled at a concentration of 5 nM and mixed with PhiX sequencing control at 5%. Libraries were sequenced on the Illumina HiSeq 4000 at a density of 12 samples per lane.

Bioinformatics

The analysis pipeline used to process sequencing results can be found under FERMI here: http://software.laliggett.com/. For a detailed understanding of each function provided by the analysis pipeline, refer directly to the software. The overall goal of the software built for this project is to analyze amplicon captured DNA that is tagged with equal length UMIs on the 5' and 3' ends of captures, and has been paired-end sequenced using dual indexes. Input fastq files are either automatically or manually combined with their paired-end sequencing partners into a single fastq file. Paired reads are combined by eliminating any base that does not match between Read1 and Read2, and concatenating this consensus read with the 5' and 3' UMIs. A barcode is then created for each consensus read from the 5' and 3' UMIs and the first five bases at the 5' end of the consensus. All consensus sequences are then binned together by their unique barcodes. The threshold for barcode mismatch can be specified when running the software, and for all data shown in this manuscript one mismatched base was allowed for a sequence to still count as the same barcode. Bins are then collapsed into a single consensus read by first removing the 5' and 3' UMIs.

Following UMI removal, consensus sequences are derived by incorporating the most commonly observed nucleotide at each position, so long as the same nucleotide is observed in at least a specified percent of supporting reads (55% of reads was used for results in this manuscript) and there are least some minimum number of reads supporting a capture (5 supporting reads was used for results in this manuscript). Any nucleotide that does not meet the minimum threshold for read support is not added to the consensus read, and alignment is attempted with an unknown base at that position. From this set of consensus reads, experimental quality measurements are made, such as total captures, total sequencing reads, average capture coverage, and estimated error rates.

Derived consensus reads are then aligned to the specified reference genome using Burrows-Wheeler (*58*), and indexed using SAMtools (*59*). For this manuscript consensus reads were aligned to the human reference genome hg19 (*60*, *61*) (though the software should be compatible with other reference genomes). Sequencing alignments are then used to call variants using the Bayesian haplotype-based variant detector, FreeBayes (*62*). Identified variants are then decomposed and block decomposed using the variant toolset vt (*63*). Variants are then filtered to eliminate any that have been identified outside of probed genomic regions. If necessary variants can also be eliminated if below certain coverage or observation thresholds such that variants must be independently observed multiple times in different captures to be included. For this manuscript, we included all variants that passed previous filters and did not eliminate those that were observed only within a single capture, unless otherwise indicated.

Elimination of false positive signal

A number of steps have been included within sample preparation and bioinformatics analysis specifically to distinguish between true positive signal and false positive signal. Using the dilution series shown in Figs. 1C-D, we can show sufficient sensitivity to identify signal diluted to levels as rare as 10^{-4}. While these dilutions show significantly improved sensitivity over many current sequencing methods, they do not address our background error rate. Unfortunately, because both endogenous and exogenous DNA synthesis is error prone, it is challenging to find negative controls that can be used to estimate background error rates with a method of mutation detection as sensitive as FERMI. Nevertheless, we have a number of steps that should eliminate most sources of false signal. The two largest sources of erroneous mutation when sequencing DNA will typically be from PCR amplification mutations (caused both by polymerase errors and exogenous insults like oxidative damage), and sequencing errors.

The steps are the following:

- *Elimination of first round PCR amplification errors*
- *Elimination of subsequent PCR amplification errors*
- *Elimination of sequencing errors*

Elimination of first round PCR amplification errors

The first round of PCR amplification performed during library preparation causes mutations that are challenging to distinguish from those that occurred endogenously. Since there is little difference between those mutations that occur during the first round of PCR amplification and those that occurred endogenously, we rely on probability to eliminate these errors. Since we are performing sequencing of individually captured alleles, we can ask whether requiring that a mutation be observed in multiple captured alleles before it is called as a true positive signal alters the frequency of variants identified. We expect about 400 first round PCR amplification errors, and the probability that the identical mutation will occur in multiple cells becomes exponentially unlikely (Fig. S1). By requiring a mutation be observed in just three captures before it is called as real signal, only about 1-2 first round PCR amplification errors should make it into the final data. In contrast, when we process our data requiring from 1 to 5 independent observations of a mutation, the overall mutation spectrum does not change, apart from a loss of the most rarely observed variants. This observation led us to include all variants that were observed even once.

Elimination of subsequent PCR amplification errors

Elimination of PCR amplification errors after the first round of PCR is done using UMI collapsing (Fig. 1a). Each time a strand is amplified, the UMI will keep track of its identity. Any mutations that occur after the first round of PCR will be found on average in 25% of the reads (or fewer for subsequent rounds). This allows us to collapse each unique capture and eliminate any rarely observed variants (<55%) associated with a given UMI. Utilizing the UMI in this way allows us to essentially eliminate any PCR amplification errors that occurred after the first round of PCR. The method should also eliminate most errors resulting from DNA oxidation in vitro.

Elimination of sequencing errors

Sequencing errors are eliminated in two ways. This first method is by using paired-end sequencing to read each strand of a DNA fragment (Fig. 1a). The sequence of these reads (Read1 and Read2) should match if no sequencing errors have been made. For an error to escape elimination it would need to occur at the same position (changing to the same new base) within both Read1 and Read2. Therefore, when the base call differs at a position on Reads 1 and 2, these changes are eliminated from the final sequence. This collapsing should eliminate most sequencing errors, although sequencing errors of the same identity occurring at the same position will escape.

These errors should be removed when collapsing into single capture bins (Fig. 1a). As with the logic when eliminating subsequent PCR amplification errors, most sequences associated with each UMI pair should be identical. Therefore, sequencing errors passing through Read1 and Read2 will be very unlikely to match other sequenced strands from the same capture event, and are eliminated during consensus sequence derivation.

Mutation signature analysis

Twenty somatic mutation signatures were previously identified (*43*) by analyzing trinucleotide mutation context of cancer genomes using non-negative matrix factorization (NMF) and principal component analysis (PCA). Here, we used deconstructSig (*64*) to identify the relative presence of those mutation signatures within the somatic mutations detected blood using somaticSignatures (*65*). Codon triplet biases were partially analyzed using the MutationalPatterns R package (*66*).

Estimation of mutation burden

It is difficult to understand the somatic lineage development that gave rise to the number of cells that are assayed from each blood biopsy. Therefore, estimating a somatic mutation rate (per cell division) is challenging. Nevertheless, we can derive estimates of somatic mutation burden.

An upper bound for the somatic mutation burden observed by FERMI analysis can be estimated by using the number of captures and total observed variants, and assume that all of these are de-novo mutations. In our data from Cohort 1, we observe on average 1,232,458 unique captures per analyzed blood sample. These captures are relatively uniformly spread across each of our 32 different probes, which span a total of 4838bp. From this, the total probed DNA, D_T, can be estimated as:

$$D_T = \frac{1232458\ captures * 4838\ bp}{32\ probes}$$

$$D_T = 186332243.9\ bp$$

The total number of observed variants within each blood sample is on average 168,940, from which the aggregate mutation burden, M, can be estimated as:

$$M = \frac{168940\ mutations}{186332243.9\ bp}$$

$$M = 9 * 10^{-4} mut/bp$$

$$M = 900\ mut/Mb$$

A lower estimate can be made by assuming that mutations are not all unique occurrences but might be the result of clonal expansions creating multiple copies of each unique mutation. This mutation burden, M, can be estimated by the approximately 40,000 captures per each of the 32 probes that captured roughly 6000 variants across a conservative 100bp sized capture for each probe (probe region is realistically smaller than 150bp because of collapsing conditions). Given that all variants for which allelic information could be discerned were present on both alleles, we can realistically conclude each of the ~3000 base positions queried was mutated at least twice (hence the estimate of 6000 variants).

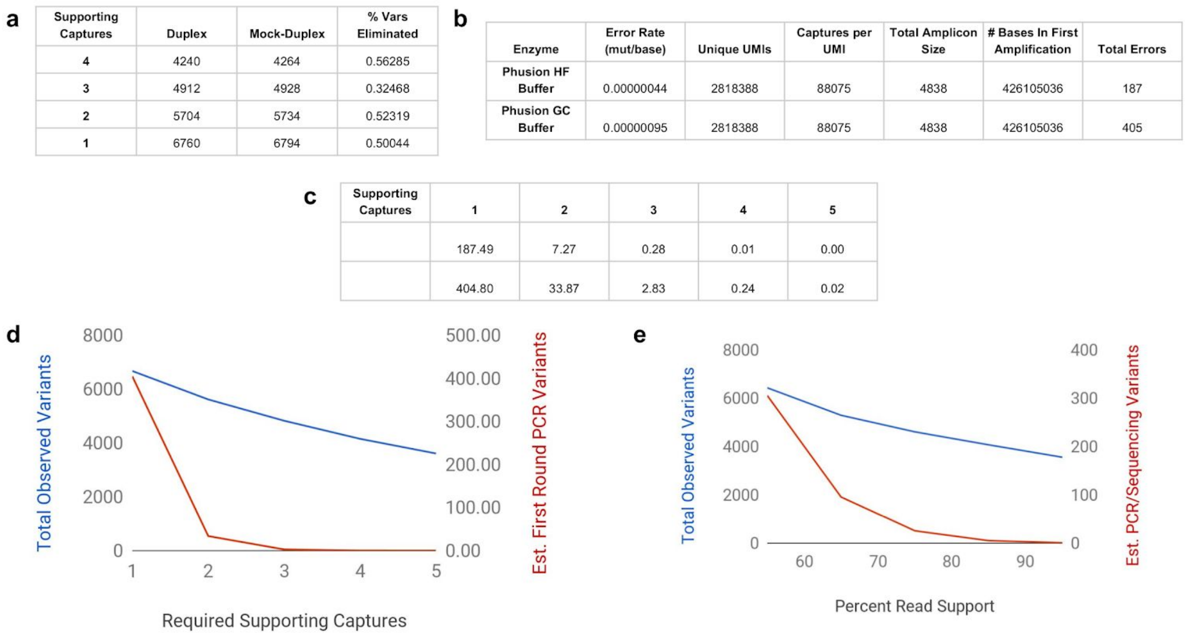
$$M = \frac{6000\ variants/sample}{40000\ captures * 32\ probes * 100\ bp/probe}$$

$$M = 5 * 10^{-5} mut/bp$$

$$M = 50\ mut/Mb$$

Supplementary Text:

Supplementary Fig. 1



a

Supporting Captures	Duplex	Mock-Duplex	% Vars Eliminated
4	4240	4264	0.56285
3	4912	4928	0.32468
2	5704	5734	0.52319
1	6760	6794	0.50044

b

Enzyme	Error Rate (mut/base)	Unique UMIs	Captures per UMI	Total Amplicon Size	# Bases In First Amplification	Total Errors
Phusion HF Buffer	0.00000044	2818388	88075	4838	426105036	187
Phusion GC Buffer	0.00000095	2818388	88075	4838	426105036	405

c

Supporting Captures	1	2	3	4	5
	187.49	7.27	0.28	0.01	0.00
	404.80	33.87	2.83	0.24	0.02

Supplementary Figure 1

Estimation of false-positive rates due to sequencing and PCR errors.

a, The use of sequencing information found within Read 1 and Read 2 of paired-end sequencing is often used to correct sequencing errors. We performed paired-end collapsing prior to consensus read derivation (Fig. 1a), though the effect was surprisingly mild. In this table, the number of identified variants are shown when duplex collapsing is used or not in consensus read derivation (mock duplexing processes the collapsing in the exact same way as duplex collapsing without eliminating variants for not being in both reads). These variant counts are shown while also varying the number of required independent supporting captures for a variant to pass filtering. The logic behind this analysis is that the fewer captures in which a variant is found, the less confidence we have that it represents true biological signal. Lower confidence variants should be more likely to be eliminated by duplex collapsing reads, if other filters were otherwise insufficient. We show that whether reads are first duplex collapsed or not, there is little effect on the percent of variants that are eliminated, suggesting that our other filtering parameters appear to adequately eliminate sequencing errors. **b**, While the filters used for FERMI should eliminate the majority of errors introduced during PCR amplification and those errors arising from sequencing mistakes, errors made in the first round of PCR amplification could be identified as false positives. If there is a sufficient number of PCR errors made within the first round of amplification, these errors could create artificial patterns within the data. Using one supporting capture as the lower limit for variants to be identified as true signal, the expected number of errors were estimated from amplification using Phusion polymerase and are shown in the table (two estimations are included because Illumina's reaction mixtures are proprietary and we do not know the exact reaction conditions). **c**, When only requiring one supporting capture, 3-6% of variants should be derived from first round PCR errors, although more than half of these will be eliminated by the requirement that 55% of reads for a capture support the variant (errors from subsequent PCR rounds will be even more efficiently eliminated by the 55% cutoff). If we require that the same variant be present at the same location across multiple captures before it is included in the final results, it becomes exponentially more unlikely that a first round PCR error would get included. In contrast, increased capture number requirements have a much more modest effect on variants called. **d**, While increasing the number of required supporting captures eliminates rare variants as well as first round PCR errors, the numbers of identified variants only decreases modestly for all individuals (blue line, left y-axis). In contrast, the number of variants expected to be identified as a result of first round PCR amplification errors exponentially decreases with each extra capture requirement (red line, right y-axis). When compared to the number of variants that pass all filters and processing, the first round PCR errors appear to have minimal effect even when only a single capture is required. Expectedly, as we increase the number of required captures supporting a variant, the total number of variants also decreases, and after two required captures should essentially not include mutations created by PCR amplification. Throughout most of this paper, a single capture is used, so as to not bias results by variant representation. Nonetheless, the patterns of mutations identified look very similar when greater numbers of supporting captures are required. **e**, As shown in Fig. 1a, when deriving consensus reads, variants are eliminated for being rarely observed across reads supporting a given capture. The cutoff we use throughout most of this manuscript is 55%, such that a given variant must be present in at least 55 percent of sequencing reads supporting a capture or they are ignored. The logic behind this chosen cutoff is that more stringent cutoffs largely do not alter the observed mutation spectra, but result in a significant loss in putatively true positive signal. With this cutoff, the expected number of sequencing errors can be estimated. We observe that 9 percent of bases are mismatched within reads supporting a given capture. Each capture is approximately 150bp in length and is supported by an average 13.5 reads. This yields an average of 182.25 errors within each sequenced capture.

$$E_{tot} = 0.09 * 150\ bp * 13.5\ reads$$

$$E_{tot} = 182.25$$

Applying the requirements that 55-95 percent of reads must support a given variant (shown as m), the number of false positive signals that pass filtering for each prepared blood sample can be computed. Within each capture there are approximately 450 total possible changes, and an average of 18 reads supporting each capture:

$$E_{seq} = m * 18\ reads/capture)^{\frac{182.5\ PCR\ err}{450\ bp}} * 1200000\ captures/sample$$

$$m = 0.55 :\ E_{seq} = 155.95\ errors/sample$$

$$m = 0.65 :\ E_{seq} = 31.48\ errors/sample$$

$$m = 0.75 :\ E_{seq} = 6.19\ errors/sample$$

$$m = 0.85 :\ E_{seq} = 1.22\ errors/sample$$

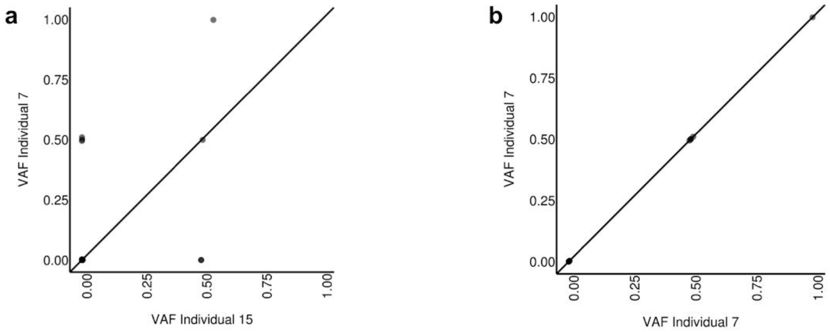
$$m = 0.95 :\ E_{seq} = 0.24\ errors/sample$$

The number of expected PCR amplification errors to pass all cutoffs is then estimated using a Gaussian distribution. The logic is that the first round of PCR amplification will create errors that will be at an allele frequency near 50 percent as an error will be created in one of two strands of a captured sequence. Using a Gaussian distribution with a mean at 50, the number of all PCR amplification errors expected to pass the 1 supporting capture and 55-95 percent of sequencing reads criteria can be calculated by integrating under the Gaussian distribution. Since we expected about 405 first round PCR amplification errors, and subsequent errors will exist at much smaller allele frequencies, the expected number of variants expected to pass criteria is calculated as follows:

$$E_{tot} = 405 * \int_{c}^{100} f(x) + m_c$$

Above we integrate from the support allele frequency *c* to *100* under the Gaussian distribution *f(x)*, multiply this by the expected total number of first round PCR amplification errors, and add to this the number of expected sequencing errors *m* as a function of the support frequency *c*. As shown here, when variants must be supported by at least one unique capture and at least 55 percent of supporting reads, we anticipate only about 150 total variants false variants to make through all FERMI analysis. We believed this to be an acceptable amount of noise given that we see about 6000 total variants from each sample and generated most of the data in this manuscript with these criteria.

Supplementary Fig. 2



c

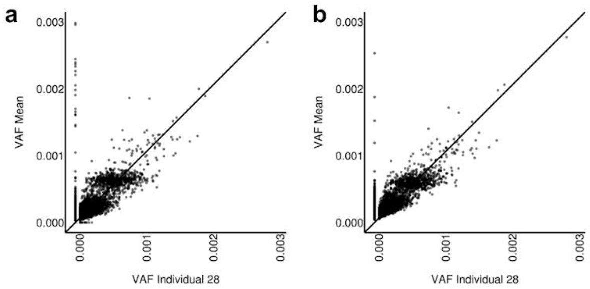
Individual	0mo vs 1mo
Individual A	0.460348
Individual B	0.538478
Individual C	0.436766
Individual D	0.522387
Individual E	0.519219
Individual F	0.482805

Supplementary Figure 2

Resequenced samples are not more similar to each other than to other individuals.

To understand how somatic mutation loads change with time, individuals were biopsied twice, 30 days apart. **a**, When comparing different individuals, low frequency variants tend to exist close to a y=x line, while high frequency SNPs differ. As expected, such SNPs cluster around frequencies of 0.5 and 1. **b**, When comparing the biopsies taken from the same individual variants show a high degree of similarity, both among SNPs and more rare variants. **c**, Though repeat sequencing of individuals typically results in close matches of rare variants with VAFs lower than 0.003, repeats do not more closely each other than they match the VAF population mean or any other typical sample. This suggests that the differences observed between samples is likely due to sampling differences than to real differences in individual mutation patterns.

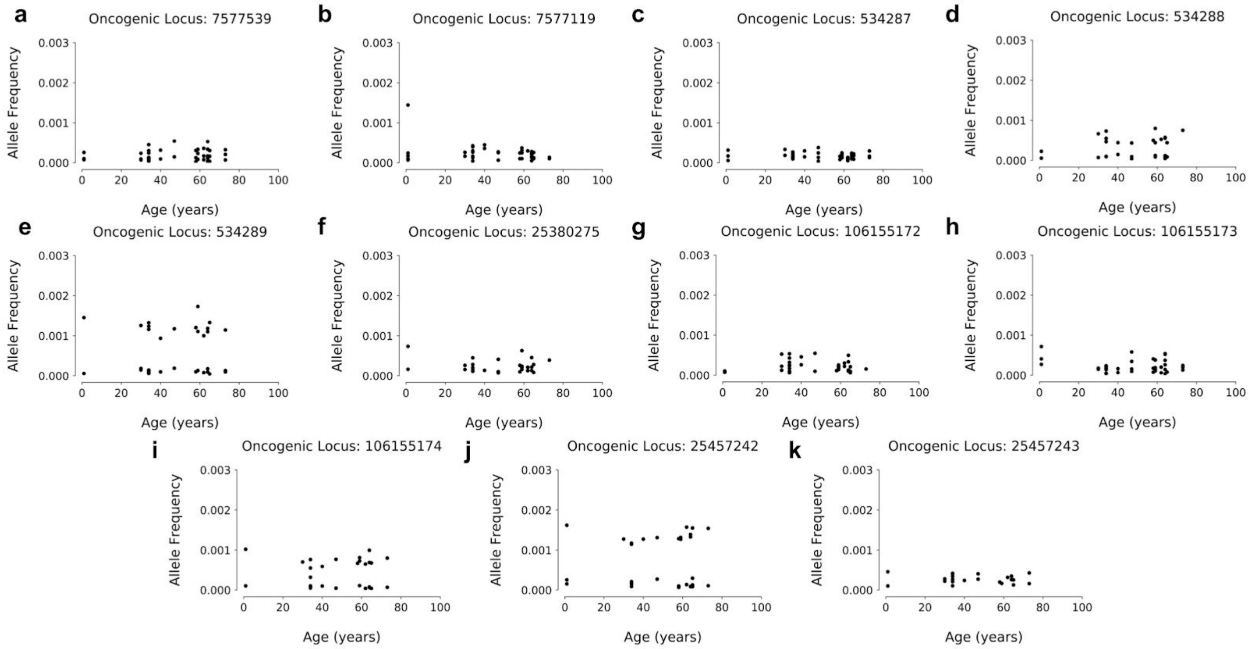
Supplementary Fig. 3



Supplementary Figure 3. Somatic mutation patterns are similar across multiple experiments.

For consistency, all samples used in the main analysis derived from a single bulk library preparation and sequencing run, forming Cohort 1. To ensure that the observed trends were not the result of some bias specific to this single preparation, the entire process was independently repeated, with eleven different blood biopsies to form Cohort 2. **a,** Individuals from Cohort 2 closely resembled mean VAFs from the Cohort 1, shown through one of the individuals from Cohort 2 (R-squared = 0.455316, p-value = 0.000000). **b**, Similarly, individuals from Cohort 2 closely resemble the overall VAF mean created from Cohort 2 samples as shown through one of the individuals from Cohort 2 (R-Squared = 0.615327, p-value = 0.000000).

Supplementary Fig. 4

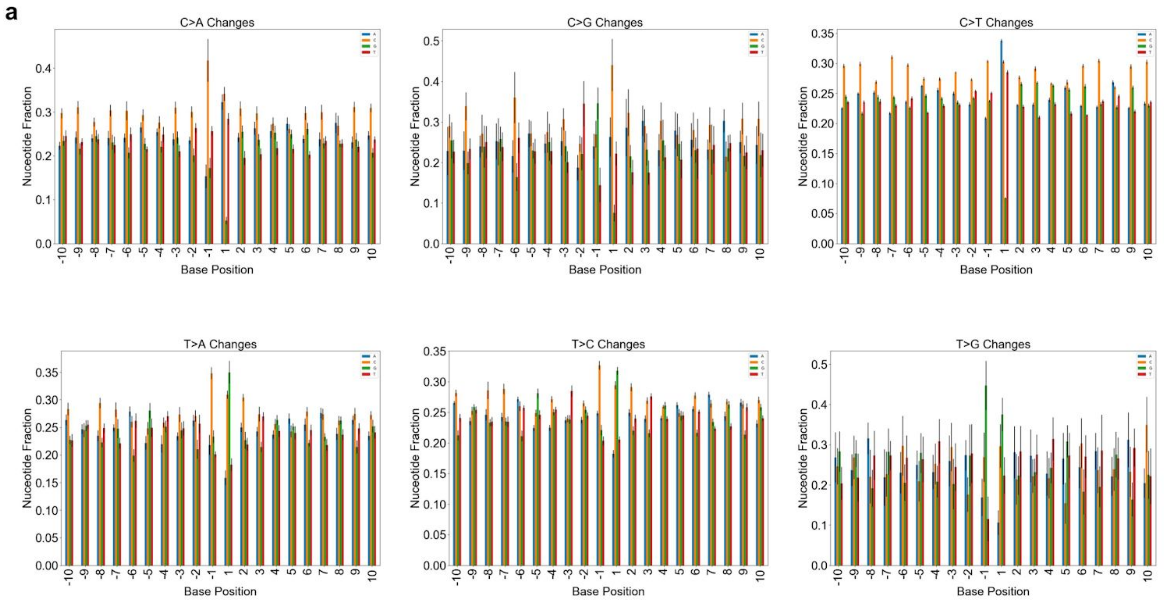


Supplementary Figure 4

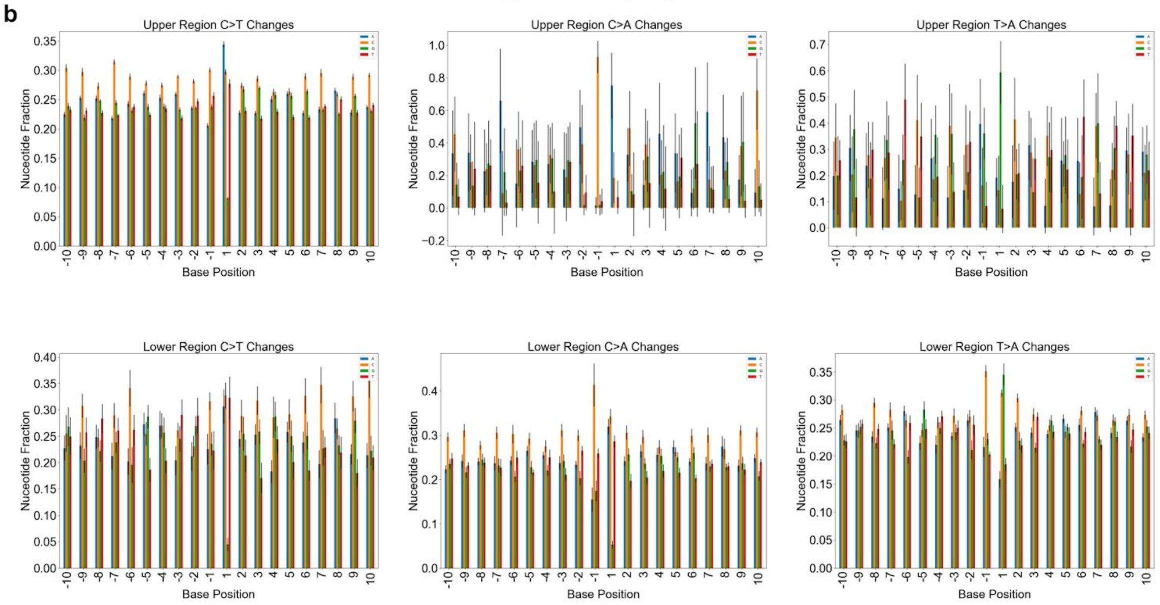
Oncogenic mutations do not show evidence of selection.

As shown in Fig. 2e, known oncogenic mutations within probed regions do not show evidence of positive selection. Shown here are additional probed oncogenic loci according the their observed VAFs across donor ages, which also do not show an increase in variant allele frequency in older ages.

Supplementary Fig. 5



Supplementary Fig. 5



Supplementary Figure 5

Trinucleotide sequence is more impactful on substitution type than further surrounding bases.

a, Shown for each of the 6 possible substitution types are the relative frequencies at which each of the various nucleotides are found at a given position surrounding the mutated nucleotide (plotted as means with standard deviation between individuals as error; n = 22). Within each of the subplots, the mutation is not shown and exists at position 0. It appears that the greatest skewing of nucleotide representation occurs at positions -1 and 1, suggesting that they have the greatest impact on how a base will mutate when it suffers a substitution. Note that for C changes, underrepresentation of

G at position 1 is expected based on low representation of CpGs in the captured regions. **b**, As seen in Fig 2a-b, substitutions tend to exist within an upper or lower region of allele frequencies. To understand if flanking nucleotide sequence plays a role in this, the populations were analyzed separately for each of 6 base changes at Cs and Ts. Suggesting the actual mutation plays a role in resulting VAF, most substitutions exist disproportionately in either the upper or lower population rather than being equally distributed between the two. For some comparisons, this resulted in larger error within one of the populations, rendering some comparisons not feasible. For C>T changes, the flanking base sequence was largely conserved between the two populations. Other substitutions show differences in flanking sequence when they exist at higher or lower VAFs, as observed for T>A at positions -1 and 1 and C>A at position 1.

Supplementary Fig. 6

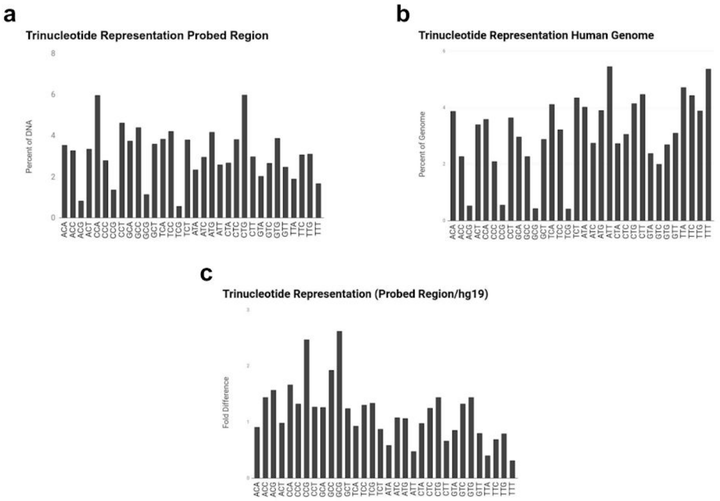
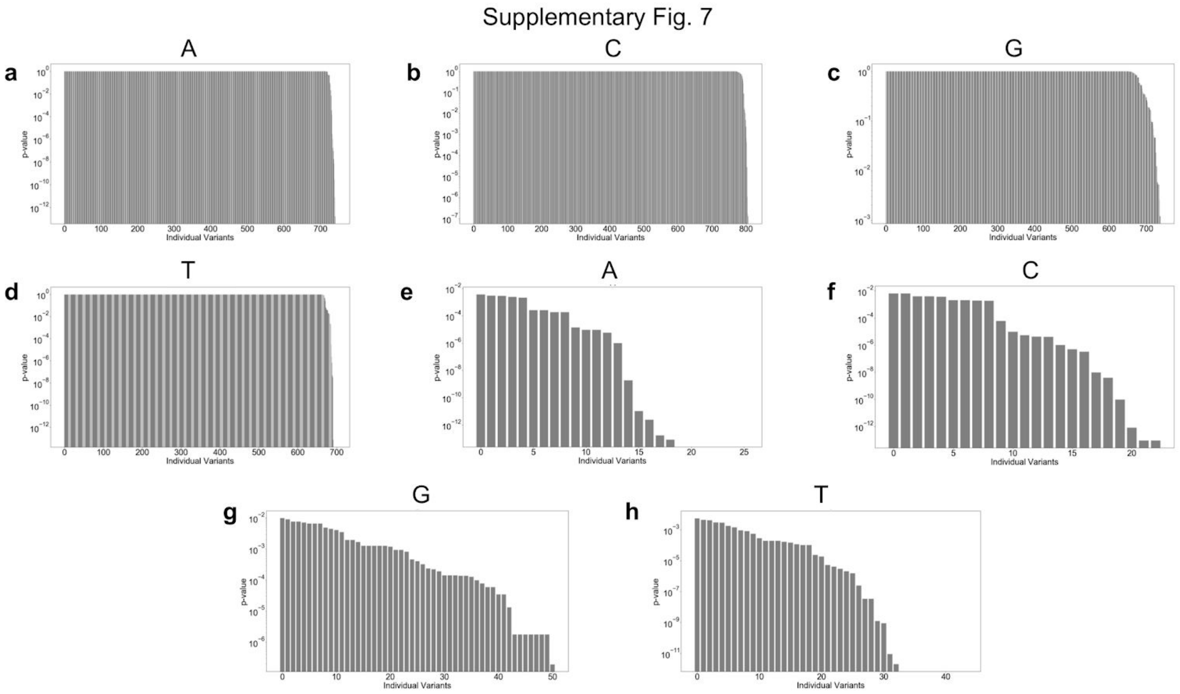


Fig. S6. Differences in nucleotide triplet representation between the probeset and human genome.

To understand how representative our total captured region was of the overall human genome, the trinucleotide sequence representation (**a**, total sequence, not identified variants) within our 32 probes was compared to the overall trinucleotide representation within the entire hg19 human genome (**b**). **c**, Comparing the trinucleotide representations within our probed region and the entire human genome demonstrates that we have a sufficient selection of CpG sites to represent the human genome as a whole.

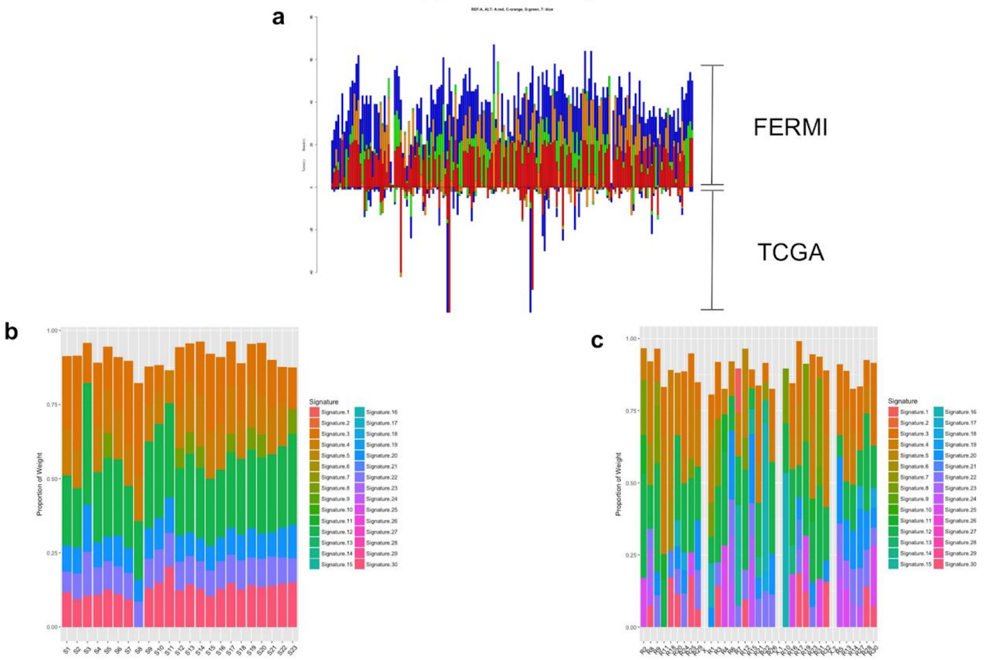


Supplementary Figure 7

Multiple positions show nonrandom base bias.

Not only is there significant conservation in the bases to which a position will change across individuals, but many locations are only observed to mutate to a single base. To understand the likelihood of this pattern arising due to chance, every instance of a given substitution was quantified for each probed site across all individuals. These changes were used to derive an overall probability that each base would change to any of the other 3 bases if mutated. Using a chi-squared algorithm to test goodness of fit, individual probabilities were computed for the base substitution pattern observed at each base locus. These probabilities were then multi-comparison corrected using Bonferroni correction, separated by reference base, ordered in descending order, and plotted here. When a variant was only observed in a small number of individuals, the probability of this change exclusively occurring at a given location due to chance was relatively high, resulting in a substantial number of non-significant loci p values ~1 (**a-d**). Plotting only positions exhibiting significant bias reveals a substantial number of bases that predictably mutate across individuals in a manner unlikely to be explained by chance (**e-h**, p values that approach zero lack bars**).** The total number of variants passing significance for each base are: A) 27 C) 23 G) 51 T) 44. These results suggest that sequence context and base location may both be playing significant roles in determining the substitution probabilities for a number of base positions throughout the genome.

Supplementary Fig. 8

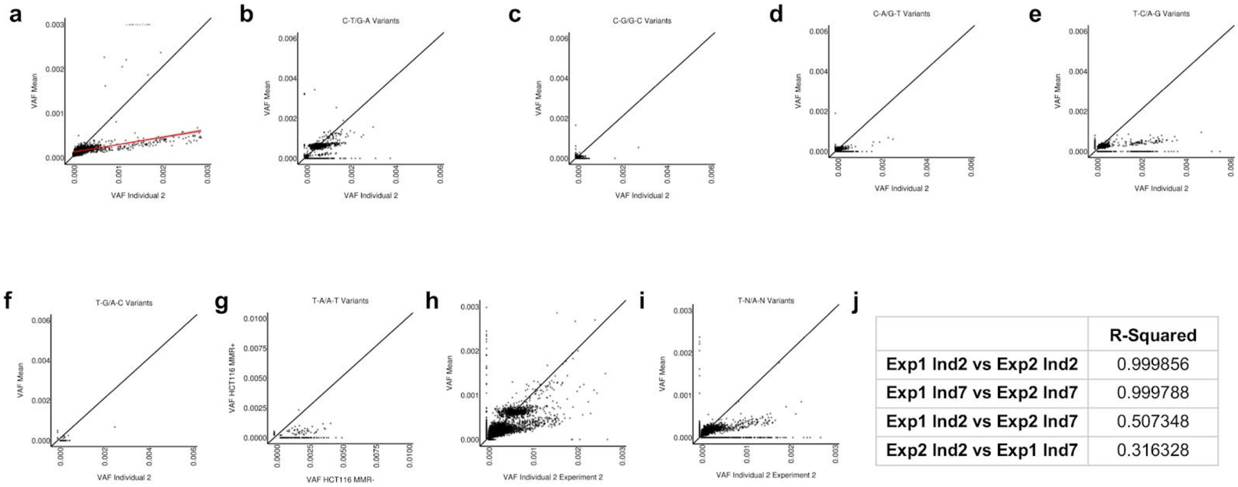


Supplementary Figure 8

Blood mutational patterns exhibit previously identified signatures distinct from those in cancers.

a, We focused on the amplicons in coding regions, and integrated Pan cancer somatic mutation data from exome sequencing in the TCGA to analyze patterns of base substitutions at genomic positions in the target regions which were mutated in both blood and tumor genomes. Substitution frequency and substitution patterns were both significantly different between blood and tumors, both at highly mutated sites (mutation count > 10; Chi square test; FDR adjusted p-value <0.05) and across all such sites (Mantel test; p-value < 1e-5), with substitution patterns in tumor genomes being more skewed. It is possible that selection during cancer evolution (as opposed to nearly neutral evolution in terminally differentiated blood cells) contribute to the observed patterns. **b**, Integrating trinucleotide contexts of the substitutions, we determined the contributions of different mutation signatures previously identified. Out of 30 previously identified signatures, our data showed overrepresentation of only 7 of them (Signatures 3, 4, 8,12, 20, 22 and 30) across different samples. Out of seven signatures, Signature 12, 3 and 4 had maximum contributions. Signature 3 and 4 are known to be associated with failure of DNA double stranded break repair by homologous repair mechanism and tobacco mutagens respectively, whereas the aetiology of Signature 12 remains unknown. **c**, There was no systematic difference in mutation signatures between amplicons when grouped by their genomic context, and they also showed similar pattern of enrichment of few signatures as compared to others, with signature 12, 3 and 4 having maximum contributions. Signature 12 and 4 exhibits transcriptional strand bias for T>C and C>A substitutions respectively, whereas signature 3 is associated with increased numbers of large InDels.

Supplementary Fig. 9



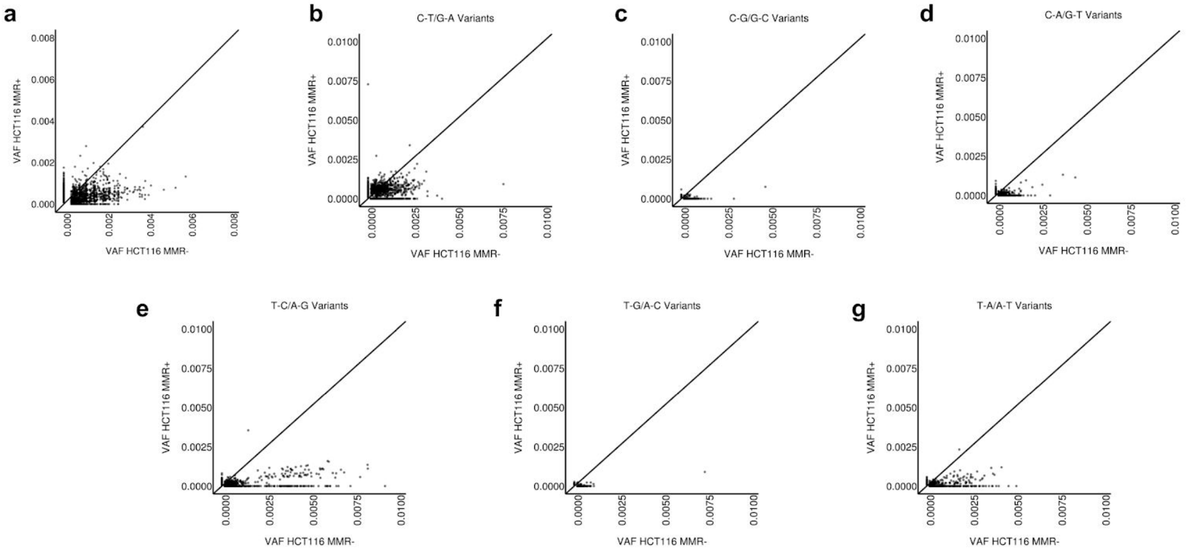
	R-Squared
Exp1 Ind2 vs Exp2 Ind2	0.999856
Exp1 Ind7 vs Exp2 Ind7	0.999788
Exp1 Ind2 vs Exp2 Ind7	0.507348
Exp2 Ind2 vs Exp1 Ind7	0.316328

Supplementary Figure 9

Base bias for cord blood individual 2 resembles MMR^{MT} Cells.

Similar to comparisons between MMR^{MT} and HCT116 parental cell lines, a cord blood donor showed a variant population that significantly deviated from expected VAFs (Fig. 6d). **a**, The mutation spectrum found within individual 2 fits to a linear regression line of y=1.9x+0.00004, from which it can be seen that variants are approximately twofold more prevalent than in the overall population average. **b-d,** Base substitutions altering C or G nucleotides did not show elevated frequencies. **e-g**, As in the in the MMR^{MT} cells, T or A changes appear at elevated frequencies. Data from individual 19 looked similar to the data shown here, but is not shown. **h-i**, To ensure that the increased frequencies of variants are not the result of experimental anomalies, the DNA from individuals #19 (not shown) and #2 was used in a second experiment. gDNA samples were freshly captured and sequenced using FERMI. In the experimental repeat, the samples showed very similar mutational spectra, with similarly elevated levels of T or A changes. For **i**, the deviating population fits a regression line of y=2.2x-9.6*10^{-5}. **j**, R^2 values are calculated (to include all variants, including germline) for the repeat sequencing of individuals 2 and 7. FERMI for the same individuals produced VAFs that were very similar between experiments, but which differed for comparisons of the two different individuals.

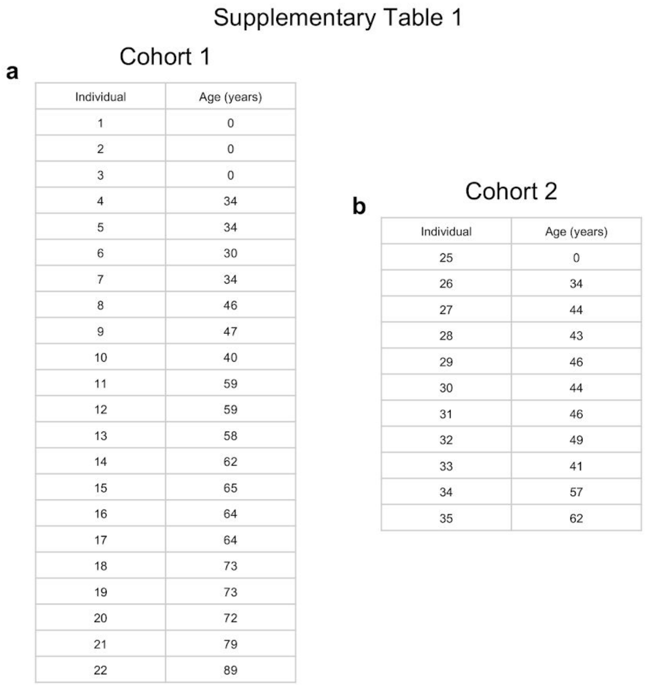
Supplementary Fig. 10



Supplementary Figure 10

MMR^{MT} VAFs are elevated over parental frequencies.

a, Verifying the observed phenotype of the MMR deficient cells in Fig. 6a, repeat sample preparation and sequencing reproduce the phenotype. Occasionally we observe slight differences in the magnitude of the increased variant allele frequencies, as evident here, but the general phenotype is always conserved. Base substitutions altering **b-d**, C or G exhibited elevated allele frequencies in MMR^{MT} cells, but substantially less compared to T or A nucleotides (**e-g**), which exhibit much higher VAFs compared to parental.



Supplementary Table 1

a Cohort 1

Individual	Age (years)
1	0
2	0
3	0
4	34
5	34
6	30
7	34
8	46
9	47
10	40
11	59
12	59
13	58
14	62
15	65
16	64
17	64
18	73
19	73
20	72
21	79
22	89

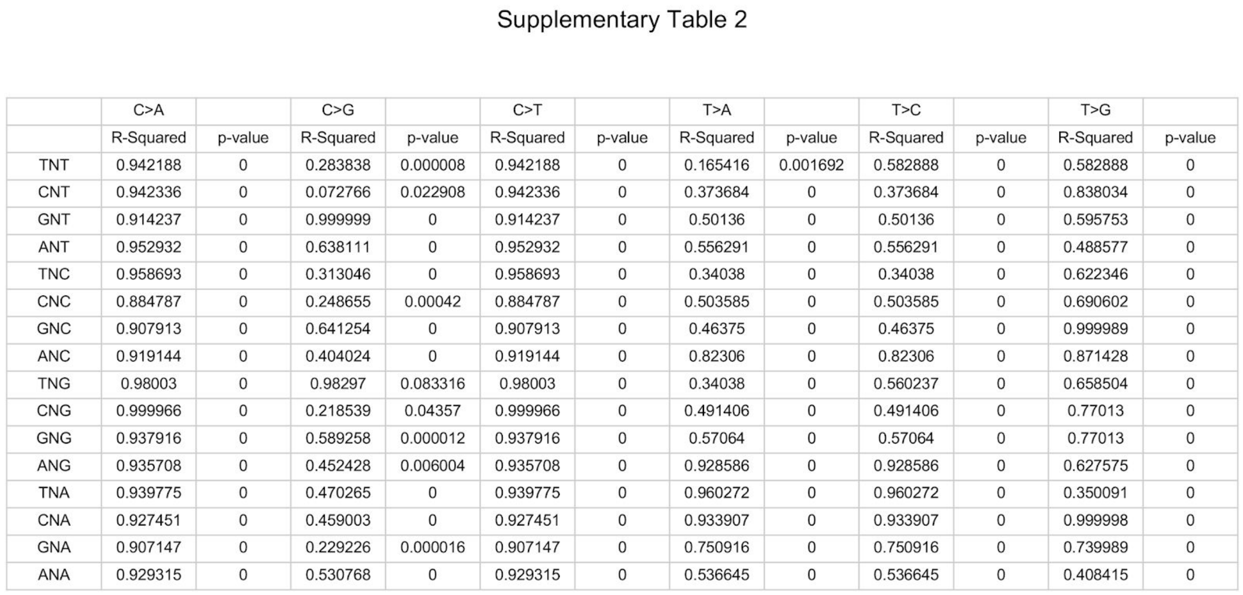
b Cohort 2

Individual	Age (years)
25	0
26	34
27	44
28	43
29	46
30	44
31	46
32	49
33	41
34	57
35	62

Supplementary Table 1

Cohort of sequenced individuals.

a, This table contains the ages of the individuals used throughout the manuscript, and their corresponding sample numbers. Those samples shown as age '0' are cord blood samples that had been previously banked. All other samples were taken from apparently healthy blood donors that passed the requirements to donate blood. **b**, This table contains the ages of individuals from a second cohort. These samples were used as the comparison to generate Extended Data Figs. 3A-B.



Supplementary Table 2

	C>A		C>G		C>T		T>A		T>C		T>G	
	R-Squared	p-value	R-Squared	p-value	R-Squared	p-value	R-Squared	p-value	R-Squared	p-value	R-Squared	p-value
TNT	0.942188	0	0.283838	0.000008	0.942188	0	0.165416	0.001692	0.582888	0	0.582888	0
CNT	0.942336	0	0.072766	0.022908	0.942336	0	0.373684	0	0.373684	0	0.838034	0
GNT	0.914237	0	0.999999	0	0.914237	0	0.50136	0	0.50136	0	0.595753	0
ANT	0.952932	0	0.638111	0	0.952932	0	0.556291	0	0.556291	0	0.488577	0
TNC	0.958693	0	0.313046	0	0.958693	0	0.34038	0	0.34038	0	0.622346	0
CNC	0.884787	0	0.248655	0.00042	0.884787	0	0.503585	0	0.503585	0	0.690602	0
GNC	0.907913	0	0.641254	0	0.907913	0	0.46375	0	0.46375	0	0.999989	0
ANC	0.919144	0	0.404024	0	0.919144	0	0.82306	0	0.82306	0	0.871428	0
TNG	0.98003	0	0.98297	0.083316	0.98003	0	0.34038	0	0.560237	0	0.658504	0
CNG	0.999966	0	0.218539	0.04357	0.999966	0	0.491406	0	0.491406	0	0.77013	0
GNG	0.937916	0	0.589258	0.000012	0.937916	0	0.57064	0	0.57064	0	0.77013	0
ANG	0.935708	0	0.452428	0.006004	0.935708	0	0.928586	0	0.928586	0	0.627575	0
TNA	0.939775	0	0.470265	0	0.939775	0	0.960272	0	0.960272	0	0.350091	0
CNA	0.927451	0	0.459003	0	0.927451	0	0.933907	0	0.933907	0	0.999998	0
GNA	0.907147	0	0.229226	0.000016	0.907147	0	0.750916	0	0.750916	0	0.739989	0
ANA	0.929315	0	0.530768	0	0.929315	0	0.536645	0	0.536645	0	0.408415	0

Supplementary Table 2

Trinucleotide context is not sufficient to predict base mutability.

To understand how well the trinucleotide context of each unique nucleotide substitution predicts base mutability, all biopsied individuals were split into two groups (Group 1 and Group 2), which were similar in ages. Within these groups, each substitution was sorted by nucleotide and trinucleotide identity. Sorted substitutions were then plotted by their VAF and compared between Group 1 and Group 2. If trinucleotide context is sufficient to predict how often and to what a nucleotide mutates, it would be expected that the comparison between Groups 1 and 2 would result in a uniformly clustered set of variants. If this were the case, the R-squared value would be small as the variant population would not fit a line. Alternatively, if factors other than just trinucleotide context were important in determining the mutability of a particular context, it would be expected that variant comparisons between Groups 1 and 2 would strongly adhere to a y=x line and therefore have a high R-squared value. For each context, the R-squared values are shown for the comparisons between Groups 1 and 2. With most comparisons showing a high R-squared value, it is clear that trinucleotide context is not sufficient to predict base mutability.