Conserved patterns of somatic mutations in human peripheral blood cells

L. Alex Liggett[1], Anchal Sharma[2], Subhajyoti De[2], and *James DeGregori[1,3,4,5]

[1]Department of Biochemistry and Molecular Genetics, [2]Rutgers Cancer Institute, New Brunswick, NJ 08901, [3]Integrated Department of Immunology, [4]Department of Pediatrics, [5]Department of Medicine, Section of Hematology, University of Colorado School of Medicine, Aurora, CO 80045

***Corresponding Author**

Corresponding Author:

James DeGregori, Ph.D.

James.DeGregori@ucdenver.edu

Summary

Mutation accumulation varies across a genome by chromosomal location, nucleotide identity, surrounding sequence, and chromatin context[1–5]. Nevertheless, while mutagens, replication machinery, and repair processes exhibit identifiable mutation signatures, at the tissue or organismal scale the aggregate somatic manifestation of these processes has been difficult to measure, and often appears to be semi-random.

This randomness is often believed to contribute to the stochasticity of diseases like cancer[6] and physiological decline during aging[7]. The challenge in observing any tissue-wide somatic mutation patterns is that prior to clonal expansion, most mutations are rare in healthy tissue[8–11]. Here we describe a new method called FERMI (Fast Extremely Rare Mutation Identification), which comprehensively captures and quantifies rare mutations at single DNA molecule resolution that exist at frequencies as rare as 10^{-4} in human peripheral blood. Using this method, we observed an unanticipated degree of ubiquity and similarity between the somatic mutation loads of different individuals, where most assayed substitutions are found to occur at conserved frequencies across nearly all individuals spanning a nine decade age range. These observed mutational patterns existed both within within non-conserved, non-coding and non-repetitive regions of the genome and within the coding regions of oncogenes implicated in hematopoietic malignancies. Furthermore, we find that nucleotides preferentially mutate to particular bases in a manner that is specified by nucleotide identity, position, and sequence context. Finally, we identify individuals who deviate from typical mutational patterns in a reproducible manner that resembles a mild mismatch repair deficiency, suggesting that variation in somatic mutation rates may be relatively common. This study provides one of the first estimates of mutation burden in terminally differentiated somatic cells and demonstrates that somatic mutations in such cells are significantly more frequent and deterministic than previously believed, and are governed by mechanisms that when perturbed, result in predictable outcomes.

Measuring somatic mutations has been technically challenging because mutations occur within individual cells that do not necessarily clonally expand to detectable representation. While these challenges have been somewhat overcome by increasing the depth of sequencing, using clever methods of barcoding[8] or by performing paired strand collapsing[12], it remains difficult to get enough sequencing depth and breadth while sufficiently limiting false positive noise[8,11,13]. To overcome these sequencing limitations, we created FERMI, in which we adapted the amplicon sequencing method of Illumina's TrueSeq Custom Amplicon platform to target only 32 x 150bp genomic regions, spanning AML-associated oncogenic mutations and the Tier III regions of the human genome (non-conserved, non-protein coding and non-repetitive). We further improved upon Illumina's capture efficiency to achieve approximately 1.2 million unique captures from 500ng-1µg of genomic DNA (gDNA) (See Methods). The targeting probes used in gDNA capture were designed with 16bp index of sequence unique to each individual and a 12bp unique molecular identifier (UMI) of random DNA, ideally unique to each capture (Fig. 1a). Sequencing reads were sorted both by sample index and UMI, producing bins of single cell sequencing which were collapsed to produce relatively error-free consensus reads. Captures were only considered if supported by at least 5 reads, and variants were only included if identified in both paired-end sequences, and detected in at least 55% percent of supporting reads (Fig. 1a and Methods; see also Extended Data Figure 1).

While all probed regions were successfully captured and amplified, capture efficiency was variable and dependent on probe identity (Fig 1b). To understand assay sensitivity, log-series dilutions of human heterozygous single nucleotide polymorphisms (SNPs) were prepared and assayed by FERMI. Using these dilutions, we observed robust quantification of diluted SNPs as rare as 10^{-4} (Fig. 1c). Even more accurate quantifications of SNP frequency can be made when using strand information to follow dilutions of multiple SNPs located on the same allele (Fig. 1d). For more description of the accuracy of FERMI, see Elimination of false positive signal in Methods and Extended Data Figure 1.

Using FERMI, we captured and sequenced gDNA from the peripheral blood of 22 apparently healthy donors ranging in age from 0 (cord blood) to 89 years of age (Extended Data Table 1). Surprisingly, within each of the probed regions, nearly every position is mutated in at least one individual, including all probed oncogenic mutations, independent of segment location or individual age, indicating a mutation rate of greater than 50 per megabase (See Estimation of mutation rate in Methods). While FERMI could correctly identify individual-specific unique germline SNPs (Extended Data Figure 2a), rare somatic variants are found at remarkably similar allele frequencies across all sampled ages. The rare allele frequencies are similar enough between most individuals that comparisons of the variant allele frequencies for each unique substitution falls along a y=x line (Fig. 2a). FERMI of biopsies taken 1 month apart from the same individuals reveal the same germline SNPs (Extended Data Figure 2b), but detected rare variants are not significantly more similar to each other than to other individuals

(Extended Data Figure 2c). Variant allele frequencies (VAFs) were averaged across 22 sampled individuals and used as a comparison to individuals, which appear age-independent and still adhere to a y=x line (R^2 Range = 0.426-0.631, Mean = 0.558) (Fig. 2b), and are similar across experiments (Extended Data Figure 3a-d shows data from an additional 11 individuals). Variants with frequencies above 0.001 were found in nearly all samples, while more rare variants were missed with a probability inversely proportional to their allele frequencies. Furthermore, most variants likely represent multiple independent events rather than clonal expansions, as they are found at similar frequencies on both alleles (Extended Data Figure 3e). It thus appears that instead of being semi-random, the aggregate effect of all DNA damage and maintenance generates somatic mutations at predictable rates throughout the genome independent of age. We suspect that such mutations primarily arise in terminally differentiated blood cell types in a sequence context-dependent manner, without constraints imposed by developmental lineages and selection, such that it reflects the basal DNA damage and repair errors in hematopoietic cells.

We observed that the overall probability of a substitution occurring is biased by nucleotide identity, with C>T substitutions being the most common and T>G substitutions being the least common (Fig. 2c). These biases were largely expected, as similar patterns have been observed both in other healthy tissues and in cancers[14–18]. There were notable differences, especially for C>N changes which we observe as underrepresented within a CpG context (Fig. 2d). Regardless of functional or oncogenic potential, each site tends to undergo the same substitutions across individuals (Fig. 2e).

These conserved substitution rates appear to be deterministic, and cannot be explained by undersampling (Extended Data Figure 4) or known base change biases (Extended Data Figure 5). It therefore appears that the combined sources of external and internal DNA mutation result in systematic substitutions at frequencies that are often predictable by location and sequence context. Suggestive of differences during cancer evolution and normal somatic mutation, the integrated exome sequencing pan cancer somatic mutation data from the TCGA exhibits different substitution patterns from those that we find in healthy donor blood (Extended Data Figure 6a). 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 (Extended Data Figure 6b-c).

While we observe variants at conserved frequencies across many individuals, previous studies have shown that selection for oncogenic changes can increase certain mutation frequencies with age[17]. While we observe each queried oncogenic change in every biopsied individual regardless of age, we do not observe significant age-related changes in the allele frequencies of either oncogenic or non-oncogenic mutations within proto-oncogenes (Fig. 2f and Extended Data Figure 7). This inability to observe any clonal expansions with age is most likely due to the fact that the average age of the individuals within the cohort is 49 years, with only 5 donors older than 70 years.

To explore the ability of FERMI to distinguish perturbations of somatic mutation patterns, gDNA from mismatch repair deficient HCT116 cells (MMR^{MT}; hemizygous for MLH1) was compared to MMR proficient parental cell line gDNA. Substantiating our

125 method, there was a substantial increase in VAFs within the MMR^{MT} gDNA when
126 compared to parental gDNA (Fig. 3a-b). Unexpectedly, while the VAFs for most
127 peripheral blood samples closely resemble those in other individuals, samples from two
128 individuals (2 and 19), contained a subset of variants that deviated from the population
129 averages with approximately a twofold increase in prevalence (Fig. 3c, 3d, and
130 Extended Data Figure 9). While the magnitude of deviation from mean VAFs was
131 different, the identities of the deviating variants were the same, such that a comparison
132 of VAFs between these two individuals correlate more closely to a y=x line than to the
133 overall population average (Fig. 3e). This consistent deviation in VAFs for these two
134 individuals from the averaged population suggests that the mechanisms governing
135 mutation levels can be systematically perturbed. Surprisingly, the VAF changes in these
136 two individuals resemble those altered in the MMR^{MT} HCT116 cells, though the
137 magnitude of these changes are greater in the latter (Fig. 3f). Finally, the deviating
138 variants found within individuals 2 and 19 are not enriched for either oncogenic variants
139 or for other variants within coding regions (Fig. 3g), indicating that deviations from the
140 typical variant pattern are not likely the result of selection.

141 As expected from previous studies[19], the HCT116 MMR^{MT} gDNA showed an
142 increased prevalence of T>C and T>A substitutions when compared to parental gDNA
143 (Extended Data Figure 8). The samples from individuals #2 and #19 also exhibited
144 these increased rates of T>C and T>A substitutions, with less extensive increases at C
145 positions, compared with the average of the 22 individuals (Fig. 3h-i and Extended Data
146 Figure 9), mirroring the changes observed in MMR^{MT} HCT116 cells. Thus, these two

individuals appear to present with a mild MMR-like substitution pattern. In support of the results, individuals #2 and #19 show the same increased rates of substitutions across multiple experiments (Extended Data Figure 9h-i). Of note, the systematic variance from the typical mutational pattern for these two individuals and the MMR^{MT} HCT116 cells serves as validation of the specificity of FERMI to accurately detect variants. More importantly, this finding of two individuals with deviating mutational patterns out of a sample size of only 22 individuals may be indicative of a broad spectrum of mutational profiles that exist across the human population.

Conclusion

These studies reveal an unprecedented degree of similarity in somatic mutational patterns across individuals, that most genomic positions are mutated within less than a million leukocytes, and how mutational spectra can be systematically disrupted in some individuals. Strikingly, we observed extremely reproducible biases at *each particular* nucleotide position in terms of the frequency of changes and the base to which it is changed. These strong position-dependent substitution biases will restrict phenotypic diversity upon which somatic evolution can act. It appears that mutation incidence, both non-oncogenic and oncogenic, are relatively well tolerated, highlighting the importance of evolved tumor suppressive and tissue maintenance mechanisms.

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.), and the Linda Crnic Institute for Down Syndrome (to J.D. and L.A.L.). The research utilized services of the Cancer Center Genomics Shared Resource, which is supported in part by NIH grant P30-CA46934.

Figure 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) to frequencies 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. **e**, ddPCR showing detection of R882H chr2:25457242 C>T mutation at approximately the expected frequencies in normal human blood.

Figure 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. **b**, VAFs from a 34 year old (x-axis) compared to mean VAFs from individuals ranging in ages from newborn to 89 years of age (n=22); R^2 =

0.590412, p=0.000. **c**, Relative contribution rates of each base substitution to all substitutions identified. **d**, Relative contribution rates of each base substitution segregated by surrounding 5' and 3' nucleotide context. **e**, All identified base substitutions within a probed region are plotted by their position and VAFs 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. **f**, Oncogenic VAFs plotted as a function of donor age does not reveal evidence of clonal expansions.

Figure 3 | 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**, MMR^{MT} vs mean VAFs from blood of the 22 individuals shows a similar pattern of increased VAFs as the comparison with parental HCT116 cells (R^2 = 0.120895). **c**, blood from a 73 yr 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). **d**, A cord blood sample (individual #2) also shows a subset of variants with higher frequencies than in the average (R^2 = 0.278250). **e**, 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). **f**, 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). **g**, Variants detected in individuals #2 and #19 are not enriched for oncogenic changes, indicated in blue. **h**, 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).

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	CCATAATTTAAAACCAAATGCTTGTGAGA232A	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	TCCTTGGCAGGTGG	GCAAGAGTGCGCTG	chr11:534258	chr1:534385	128

	GGCAGGAGACCC	ACCATCCA			
KRAS-1	AGGTACTGGTGGAG TATTTGATAGTGT	CAAGAGTGCCTTGA CGATACAGCTAATT	chr12:25398247	chr12:25398415	169
KRAS-2	GACTGTGTTTCTCC CTTCTCAGGATTC	TACAGTGCAATGAG GGACCAGTACATG	chr12:25380242	chr12:25380368	127
TET2-1	CCATGTTTTGGCTC ATTCATGCTCTTA	ACGGCCACTCCCCC AATGTCAG	chr4:106197237	chr4:106197405	169
TET2-2	CTTTTGAAAGAGTG CCACTTGGTGTCT	GGTGATGGTATCAG GAATGGACTTAGTC	chr4:106155137	chr4:106155275	139
DNMT3A	TGTGTGGTTAGACG GCTTCCGGGCA	AGGCAGAGACTGCT GGGCCGGTCA	chr2:25457211	chr2:25457364	154
IDH1	CAAATGTGGAAATC ACCAAATGGCACC	TGGGGATCAAGTAA GTCATGTTGGCA	chr2:209113077	chr2:209113239	163
IDH2	GAAGAAGATGTGGA AAAGTCCCAATGG	CATGGCGACCAGGT AGGCCAGG	chr15:90631809	chr15:90631969	161
GATA1	CTTCCAGCCATTTC TGAGATATCCTCA	CAGCTGCAGCGGT GGCTGTGCT	chrX:48649667	chrX:48649849	183
SF3B1	GTGAACATATTCTG CAGTTTGGCTGAA	ACCATCAGTGCTTT GGCCATTGC	chr2:198266803	chr2:198266967	165
TIIIA	CATCTATTCTGTGCT AGGCATTGTGTG	CAGACCTAGCATCT GTGCCAGAC	chr1:115227814	chr1:115227978	165
TIIIB	CAGTCTGGGTTTTG GAGCAATGATATC	GCAGTGAGCTCAGC CTTGATTTT	chr2:223190674	chr2:223190820	147
TIIIC	CCTGGTGCTTAGTC CTGTTCTGAAATT	AGTCTTCTATAATGC CACAACCTGTAT	chr2:229041101	chr2:229041289	189
TIIID	GAACAGAACACTTG GTAGTTGACCATG	AGACAGGGAACTGG CATGAAGAGTTT	chr4:110541172	chr4:110541302	131
TIIIE	GCCTAGAACAGGCA CCATACATTCAAT	AGATGGTGTTGCTG TGCCGGATAGGAG	chr4:112997214	chr4:112997386	173
TIIIF	TGGCACTATGTGGA GATGTTAGTACAG	GGATGTTGGTGCTA TCAGTAGCCATA	chr4:121167756	chr4:121167884	129
TIIIG	CTCTAGGCTTAGTG GTCAAGGAATGAA	AGAAGCAGGACTGT GCTTCCAAACAA	chr4:123547743	chr4:123547901	159
TIIIH	CTTGGTGGTAGCCT AGGCAGTAATTAA	CACGTGGTTGGGAA GAGAAAGTG	chr4:124428637	chr4:124428767	131
TIIIJ	TTCTATAGCACTGG TGACCAGGACACT	CTGGCCACAGTGCC TGGTTTCC	chr11:2126256	chr11:2126420	165
TIIIK	AGACAGGAGGAAG GAGCAATTCAGAAG	CATGGAGATCTCGT CCCCTCAGA	chr11:2389983	chr11:2390171	189
TIIIL	TAGGCCAGAAAACA CACAGTGTCGGG	AACTCCGGTAAGTG GCGGGTGGGGGT	chr11:2593889	chr11:2594074	186
TIIIM	ATCTGGGAACAGAC CTTCCTCAGGCAT	GTTCTAAGTTACTCT GTGTACTTGACT	chr11:11486596	chr11:11486728	133
TIIIN	AGCCTAGTTACCAT AGACGGATTCAAC	GAATATCTTCTAACT GGACTTAGAAAACC	chr15:92527052	chr15:92527176	125
TIIIO	CCAACATGTTCTAA ATTCTGGCCACAG	TGGGTCTCAGCCAT CCCATTACTG	chr16:73379656	chr16:73379832	177
TIIIP	CTAACATCTCACTTC TACCCTACGCTA	TAAGTGCCCACTAC CCCATCCTTAAT	chr16:82455026	chr16:82455164	139
TIIIQ	TCATGACCCAGGCC TCCCAGAACTGAG	ATCTGTGAAGCCGG AGTGAAAACAAC	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 95ºC in a heat block (SciGene Hybex Microsample Incubator Catalog #1057-30-O). Heat block temperature was then set to 60ºC, allowed to passively cool from 95ºC and incubated for 24hr. Following incubation, the heat block was set to 40ºC 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 37ºC 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[20], and indexed using SAMtools[21]. For this manuscript consensus reads were aligned to the human reference genome hg19[22,23] (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[24]. Identified variants are then decomposed and block decomposed using the variant toolset vt[25]. 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 putatively 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 single-cell sequencing, we can require that a mutation be observed in multiple cells before it is called as a true positive signal. As we expect about 400 first round PCR amplification errors, the probability that the identical mutation will occur in multiple cells becomes exponentially unlikely (Extended

Data Figure 1). By requiring a mutation be observed in just three cells before it is called as real signal, only about 1-2 first round PCR amplification errors should ever make it into the final data. When we process our data requiring up 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. Our logic is that while about 400 variants will be the result of the first round of PCR amplification, these same variants are already occurring endogenously, meaning that absolution variant allele frequency accuracy will be affected, but not the identities of variants.

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 in 25% of the reads or fewer. This allows us to collapse each unique capture and eliminate any rarely observed variants 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.

Elimination of sequencing errors

Sequencing errors are eliminated in two ways. This first method is by using paired-end sequencing to sequence the same fragment of DNA twice (Fig. 1a). The

information found within each of these reads (Read1 and Read2) should theoretically be the same, meaning that unless the same sequencing error is made at the same locus within both Read1 and Read2, the two strands will differ. Any differences are simply eliminated from the data as it is unknown which base call is correct, and the rest of the data is included in the analysis. This collapsing should eliminate most sequencing errors, though it will sequencing errors of the same identity occurring at the same locus. These errors are removed when collapsing into single cell bins (Fig. 1a). As with the logic when eliminating subsequent PCR amplification errors, all sequence associated with each UMI should be identical. Therefore, sequencing errors passing through Read1 and Read2 will not match other sequenced strands from the same capture event, and are eliminated during consensus sequence derivation.

Mutation signature analysis

20 somatic mutation signatures were previously identified[15] by analyzing trinucleotide mutation context of cancer genomes using non-negative matrix factorization (NMF) and principal component analysis (PCA). Here, we used deconstructSig[26] to identify the relative presence of those mutation signatures within the somatic mutations detected blood using somaticSignatures[27]. Codon triplet biases were analyzed using the MutationalPatterns R package[28].

Estimation of mutation rate

It is difficult to understand the somatic lineage development that gave rise to the number of cells that are assayed from each blood biopsy. So estimating a somatic mutation rate is challenging. Nevertheless we can estimate what somatic mutation rates might be assuming the number of cells that were assayed contained all unique variants or not.

An upper bound for the somatic mutation rate 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 we observe about 1232458 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 observed variants within each blood sample is approximately 168940, from which the aggregate mutation rate, 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 many copies of each mutation. This mutation rate, M, can be roughly estimated by the approximately 40000 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).

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

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

$$M = 50\ mut/Mb$$

1. Benzer, S. ON THE TOPOGRAPHY OF THE GENETIC FINE STRUCTURE. *Proc. Natl. Acad. Sci. U. S. A.* **47,** 403–415 (1961).
2. Gaffney, D. J. & Keightley, P. D. The scale of mutational variation in the murid genome. *Genome Res.* **15,** 1086–1094 (2005).
3. Lercher, M. J., Williams, E. J. B. & Hurst, L. D. 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).
4. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
5. Hwang, D. G. & Green, P. 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).
6. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355,** 1330–1334 (2017).
7. Kujoth, G. C. *et al.* Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* **309,** 481–484 (2005).
8. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23,** 843–854 (2013).
9. Preston, J. L. *et al.* High-specificity detection of rare alleles with Paired-End Low Error Sequencing (PELE-Seq). *BMC Genomics* **17,** 464 (2016).

10. Zhang, T.-H., Wu, N. C. & Sun, R. A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC Genomics* 1–9 (2016).
11. Schmitt, M. W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods* 1–4 (2015).
12. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9,** 2586–2606 (2014).
13. Chen, L., Liu, P., Evans, T. C., Jr & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355,** 752–756 (2017).
14. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538,** 260–264 (2016).
15. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
16. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354,** 618–622 (2016).
17. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* 1–11 (2014).
18. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).
19. Zhao, H. *et al.* Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. *eLife Sciences* **3,** e02725

(2014).

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

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

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

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

24. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

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

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

27. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31,** 3673–3675 (2015).

28. Blokzijl, F., Janssen, R., Van Boxtel, R. & Cuppen, E. MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues. *bioRxiv*

071761 (2016). doi:10.1101/071761

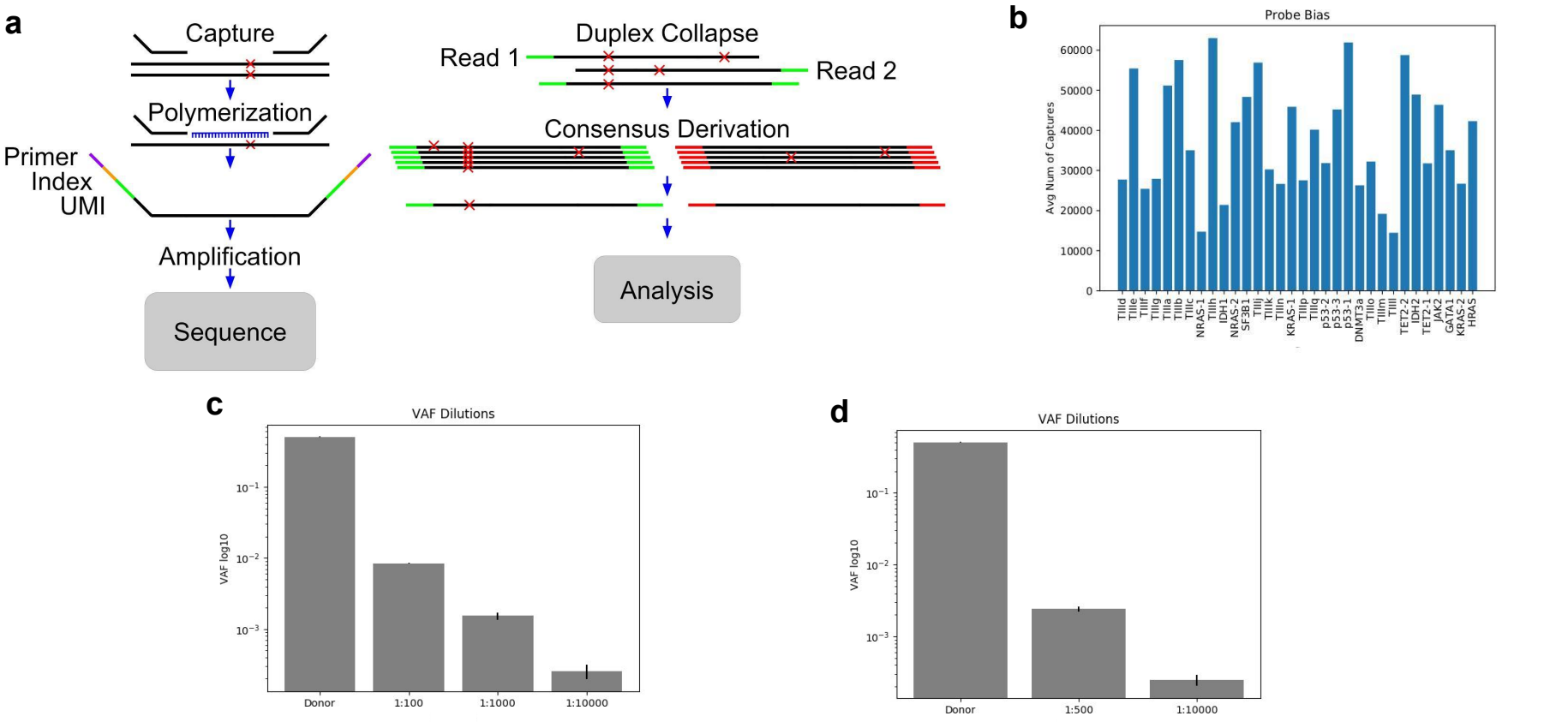


Figure 1 | Amplicon sequencing accurately detects mutation allele frequencies as rare as 1/10,000. a, Graphical depiction of gDNA capture and analysis method. **b**, Capture efficiencies vary in a probe dependent manner. **c**, Accurate detection of a single heterozygous SNP in gDNA from one individual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

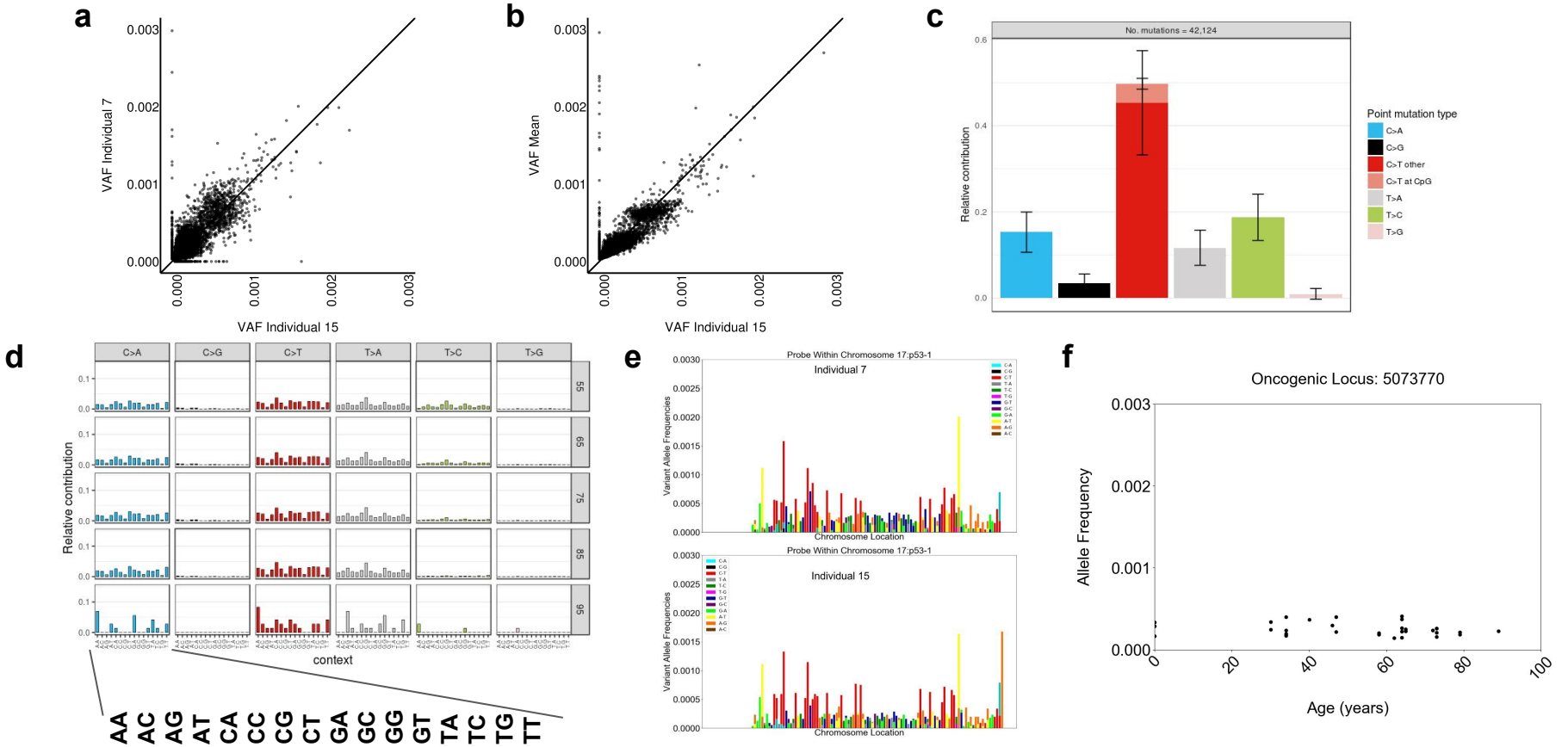


Figure 2 | Mutations exist at conserved frequencies independently of age. **a**, Comparison of VAFs of identified variants within a 34 year old (x-axis) and 62 year old (y-axis); R^2 = 0.408211, p=0.000. **b**, VAFs from a 34 year old (x-axis) compared to mean VAFs from individuals ranging in ages from newborn to 89 years of age (n=22); R-Squared = 0.590412, p=0.000. **c**, Relative contribution rates of each base substitution to all substitutions identified. **d**, Relative contribution rates of each base substitution identified by surrounding 5' and 3' nucleotide context. **e**, All identified base substitutions within a probed region are plotted by their position and VAFs for individuals 7 and 15 (representative of most other individuals), revealing highly reproducible patterns. **f**, Oncogenic VAFs plotted as a function of donor age show little evidence of clonal expansion.

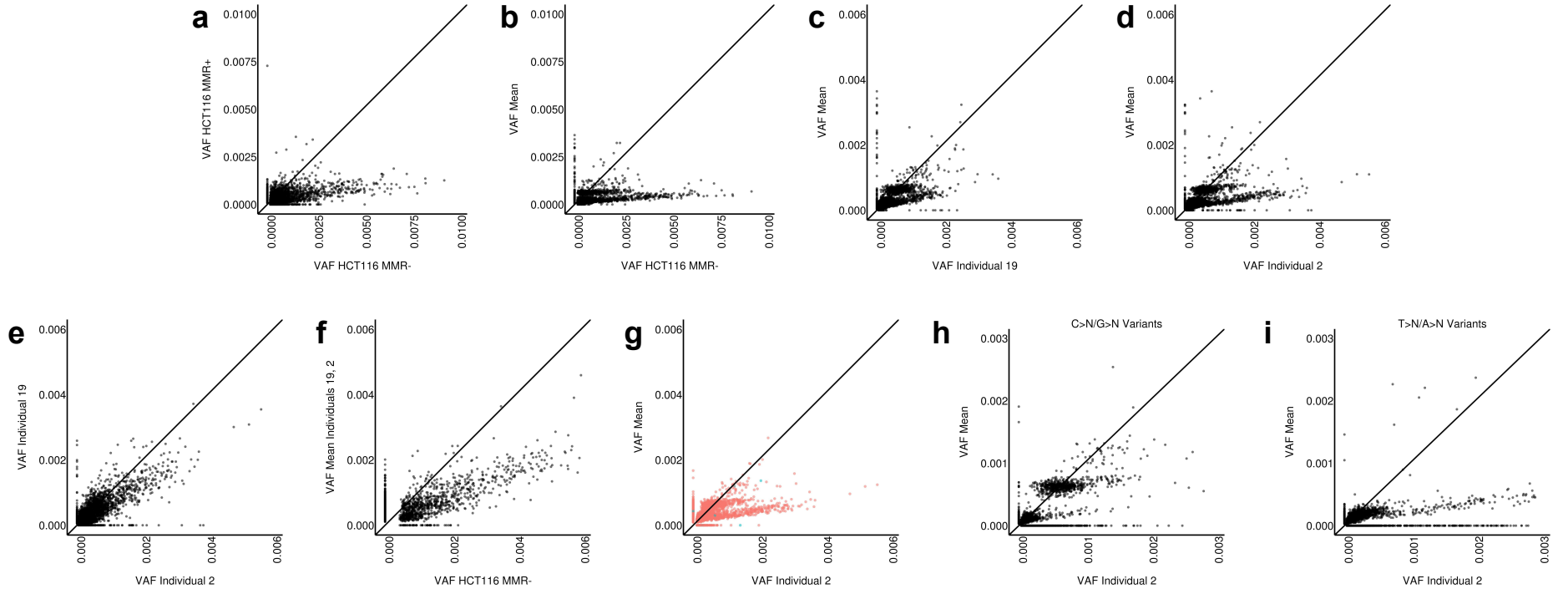


Figure 3 | Individuals Can Systematically Deviate from Population Average. a, Comparing VAFs in HCT116 MMR+ vs MMR^{MT} cells reveals an increase in frequencies for many of the observed variants in MMR^{MT} cells (R-Squared = 0.211479). **b**, MMR^{MT} vs mean VAFs from blood of the 22 individuals shows a similar pattern of increased VAFs as the comparison with parental (R-Squared = 0.120895). **c**, blood from a 73 yr 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-Squared = 0.387125). **d**, A cord blood sample (individual #2) also shows a subset of variants with higher frequencies than in the average (R-Squared = 0.278250). **e**, 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-Squared = 0.613542). **f**, 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 identities of the deviating variants are the same (R-Squared = 0.587474). **g**, Variants detected in individuals #2 and #19 are not enriched for oncogenic changes, indicated in blue **h**, Plot of only C>N/G>N variants shows relative similarity between MMR- and parental cells (R-Squared = 0.350623). **i**, Plot of only T>N/A>N variants reveals that the majority of deviating variants between MMR^{MT} and parental cells are substitutions affecting T or A.

Extended Data Figure 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

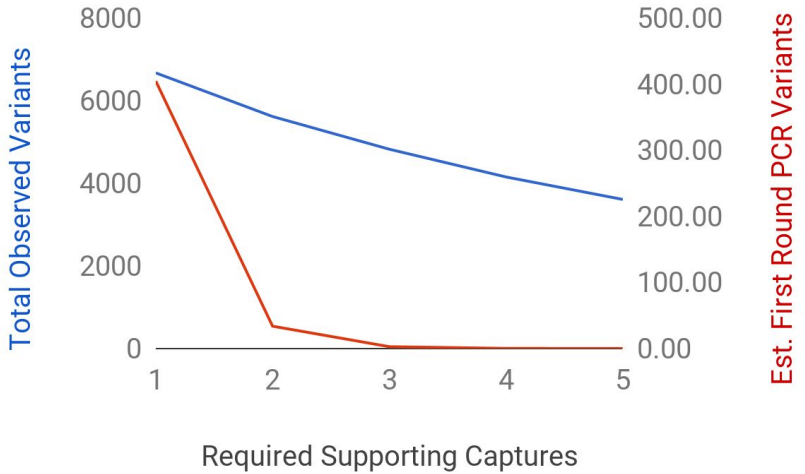
b

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

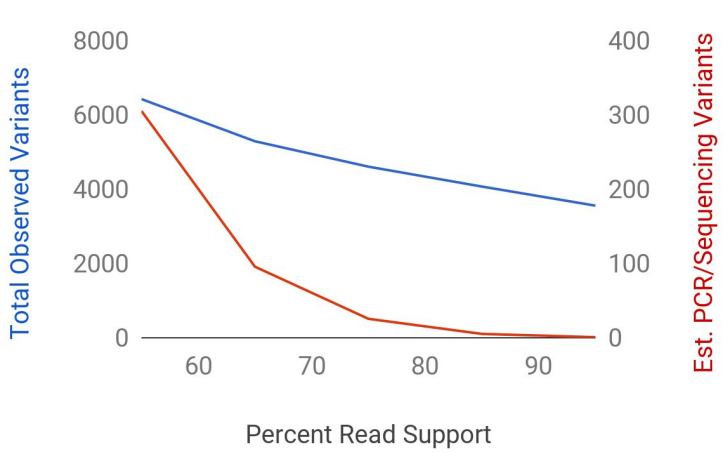
c

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

d



e



Extended Data 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

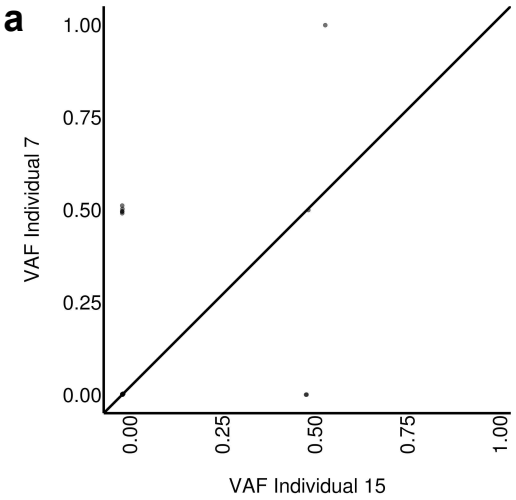
b

Cohort 2

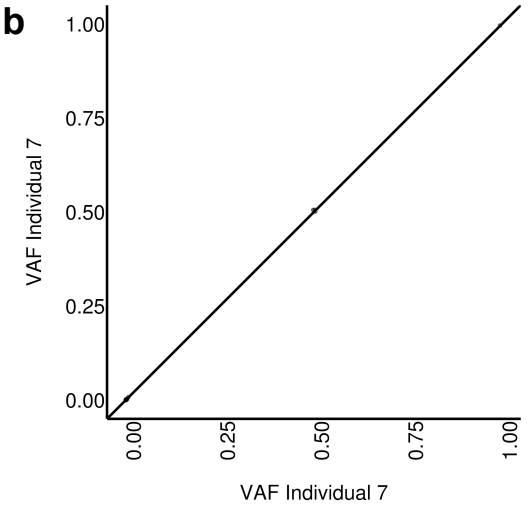
Individual	Age (years)
25	0
26	34
27	44
28	43
29	46
30	44
31	46
32	49
33	41
34	57
35	62

Extended Data Figure 2

a



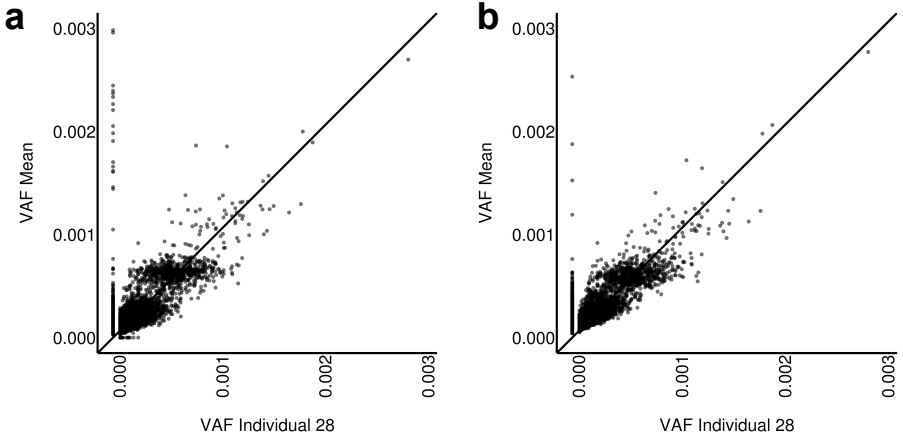
b

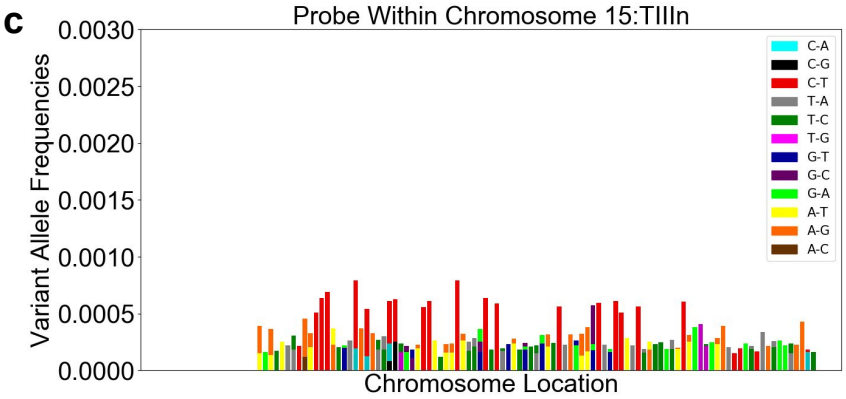


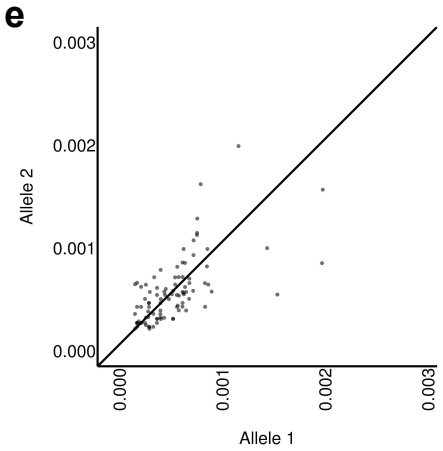
c

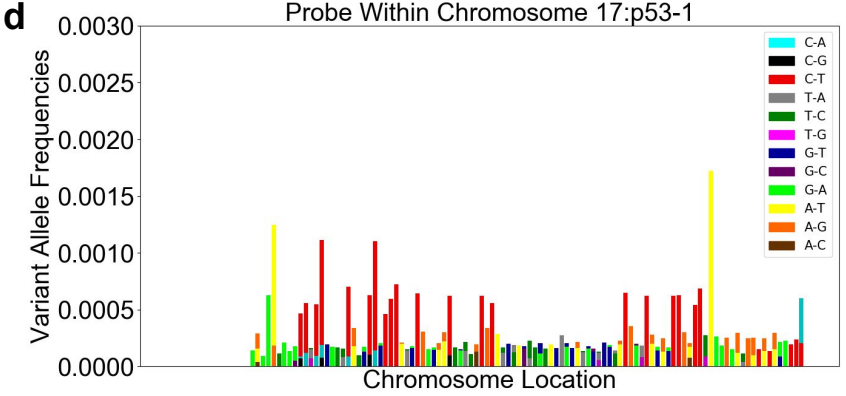
Individual	0mo vs 1mo
Individual A	0.460348
Individual B	0.538478
Individual C	0.436766
Individual D	0.522387
Individual E	0.519219
Individual F	0.482805

Extended Data Figure 3



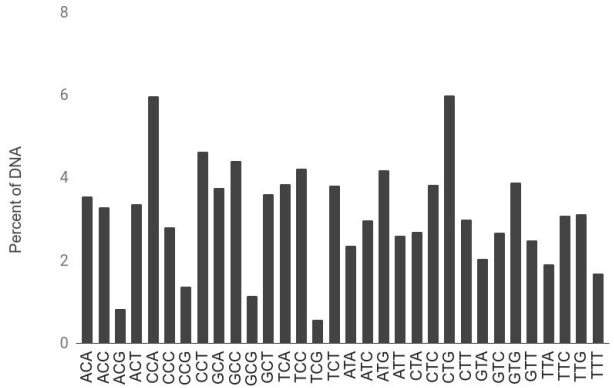




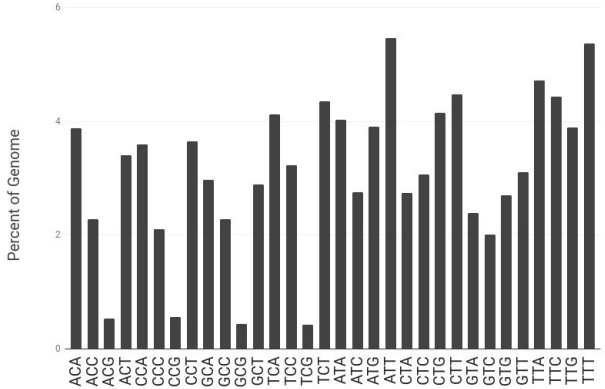


Extended Data Figure 4

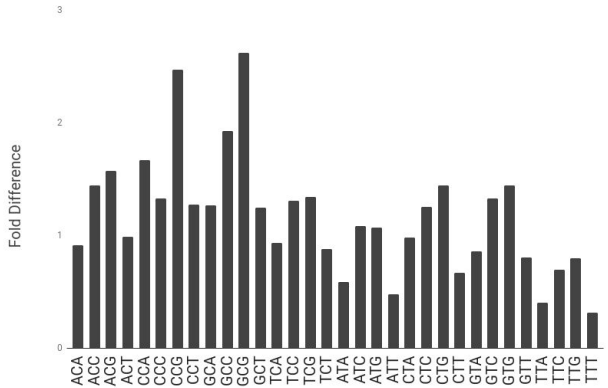
a



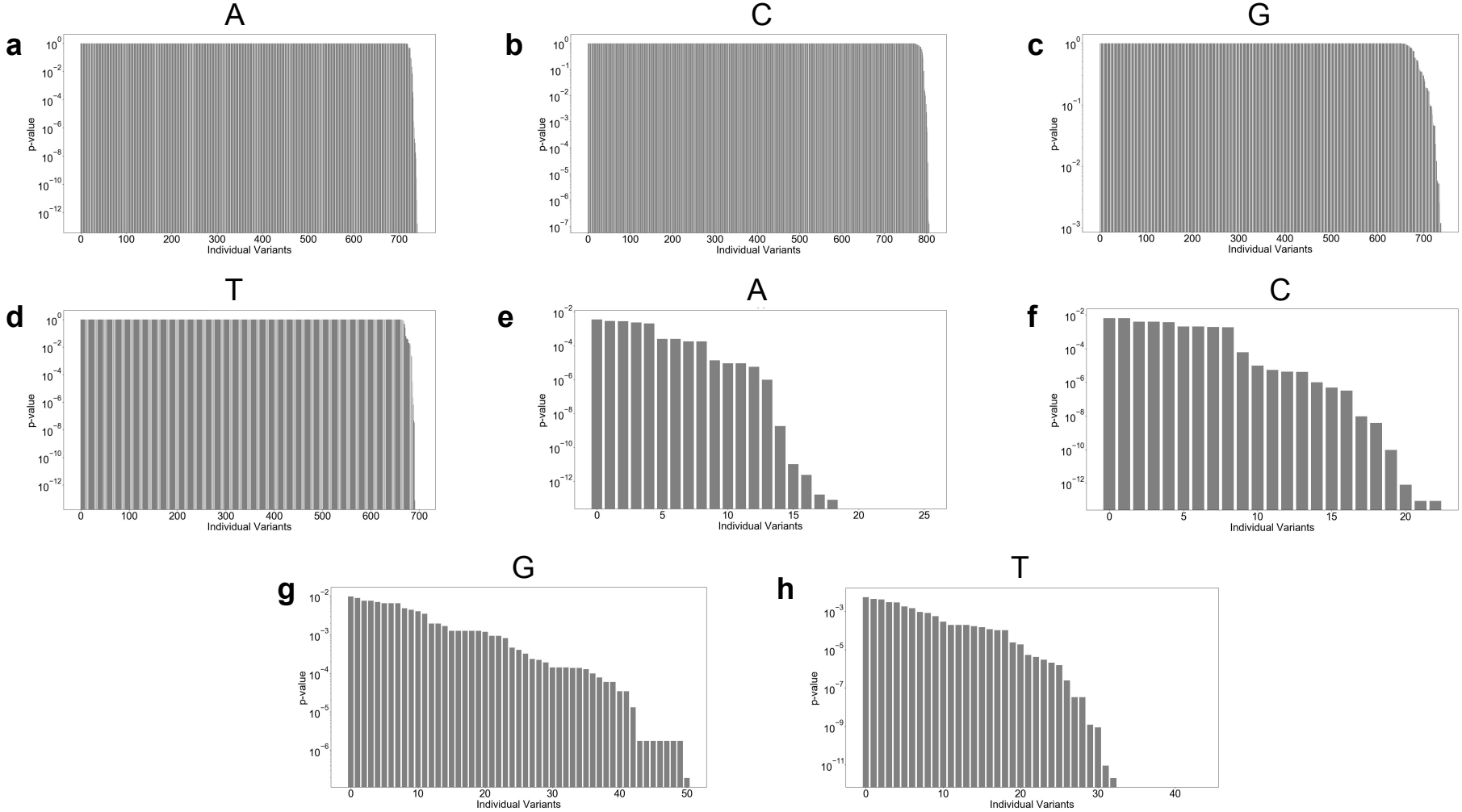
b



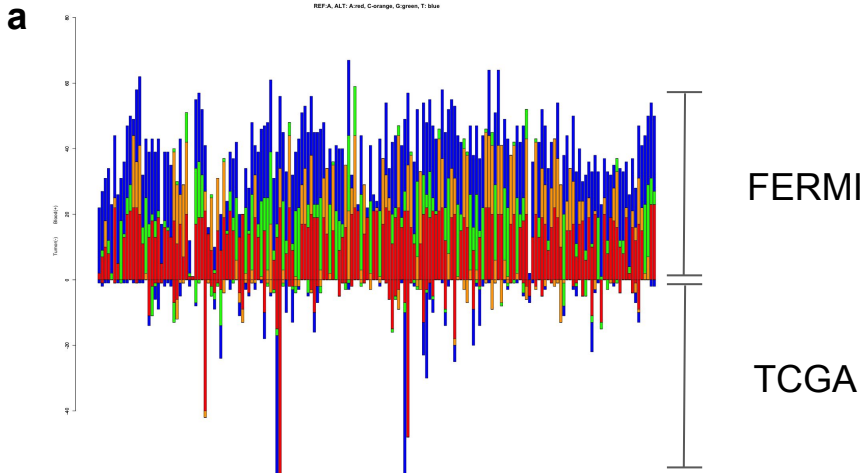
c

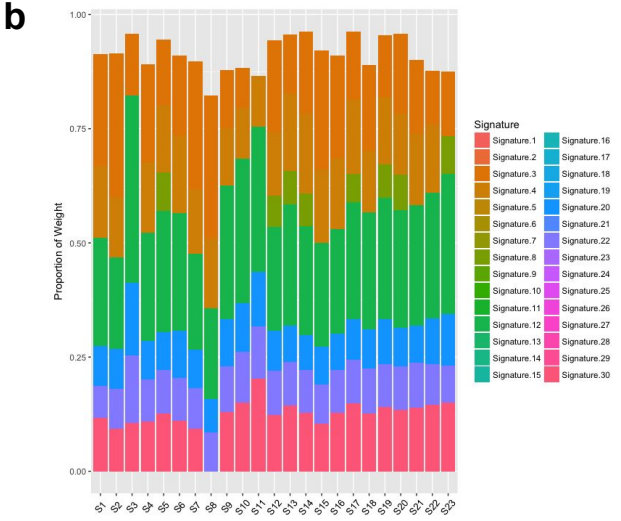


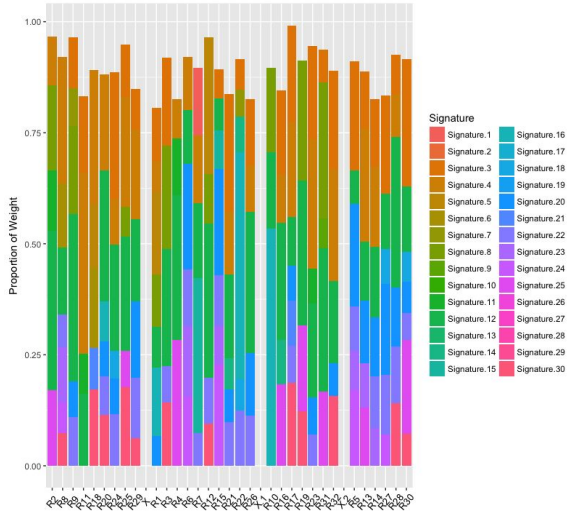
Extended Data Figure 5



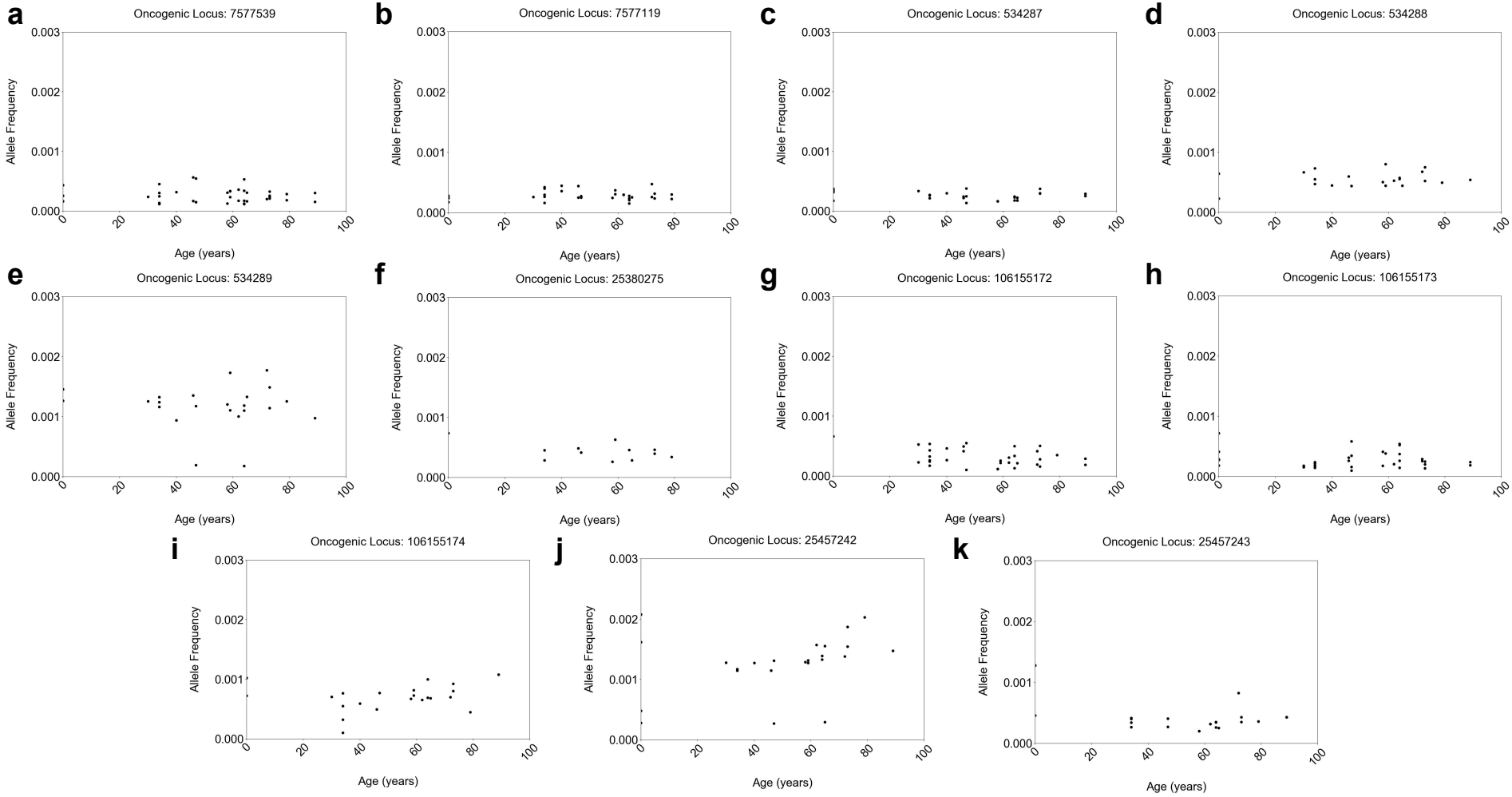
Extended Data Figure 6



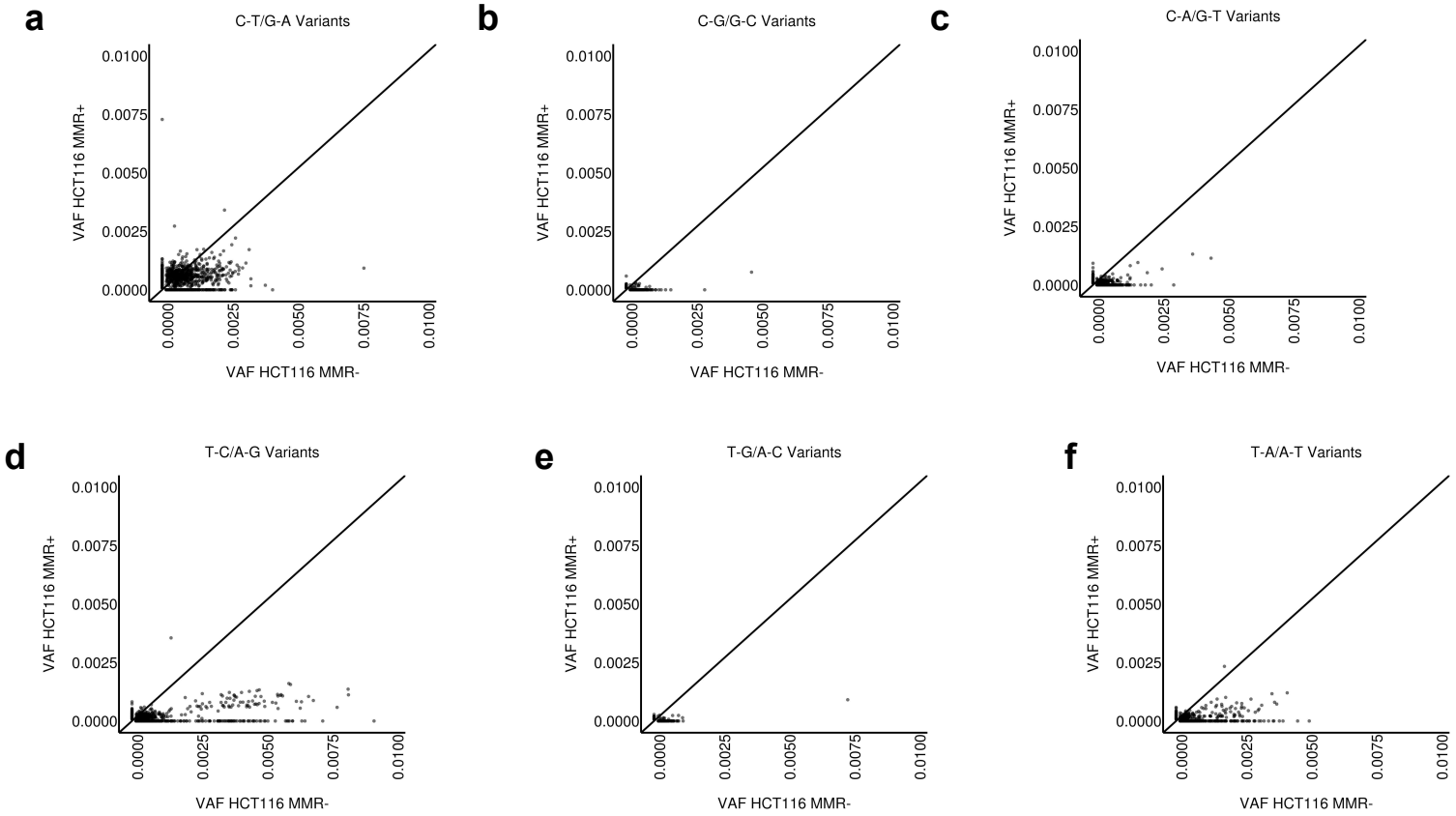




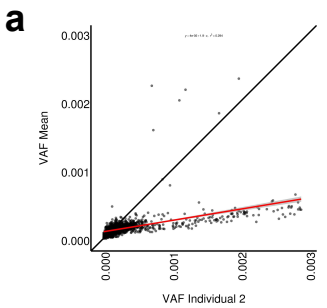
Extended Data Figure 7

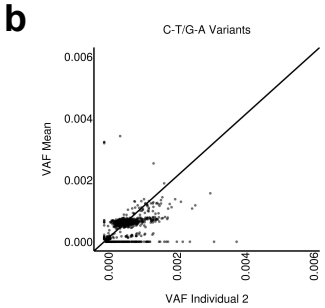


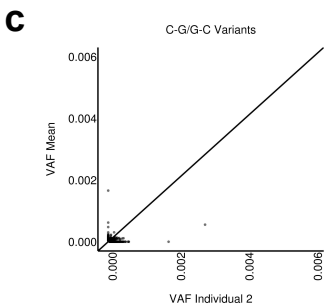
Extended Data Figure 8

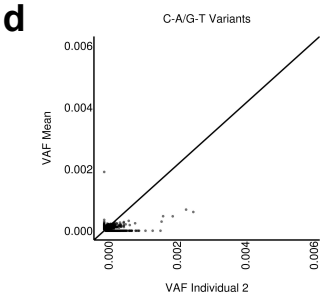


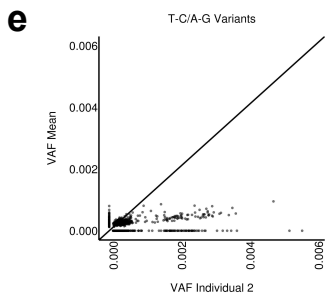
Extended Data Figure 9

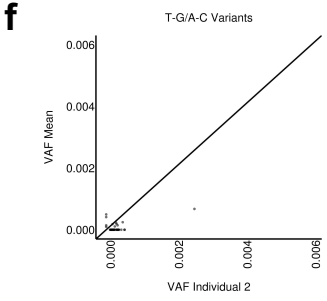


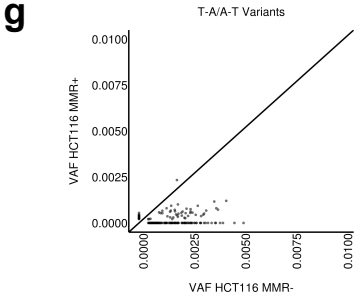


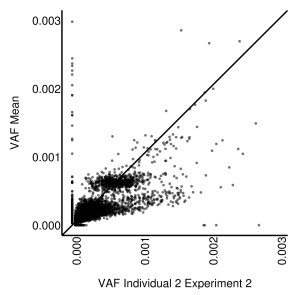












i

	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

Extended Data 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.

Extended Data 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 used to ensure that the data generated by FERMI was not experiment specific. These samples were used as the comparison to generate Extended Data Figs. 3a-b.

Extended Data Figure 2: Resequenced samples are not more similar to each other than to other individuals.

a, Low frequency variants tend to exist close to a y=x line, while high frequency SNPs differ across individuals. As expected, such SNPs cluster around frequencies of 0.5 and 1 (R-Squared=0.243364). **b**, When samples are re-sequenced, they show a high degree of similarity, both among SNPs and more rare variants (R-squared=0.568749). **c**, Though repeat sequencing of individuals typically results in close matches of VAF, 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 loads.

Extended Data Figure 3: Variants detected represent multiple independent events and reproduce across multiple experiments.

For consistency, all samples used in the main analysis derive from a single bulk library preparation and sequencing run. To ensure that the observed trends are not the result of some bias specific to this single preparation, the entire process was independently repeated, with eleven different blood biopsies (Cohort 2). **a**, Cohort 2 samples closely resembled averaged allele frequencies from the Cohort 1 (R-squared = 0.455316, p-value = 0.000000). **b**, Comparing Cohort 2 samples against the VAF mean created from Cohort 2 samples produces a similar pattern to the same comparison using the Cohort 1 data (R-Squared = 0.615327, p-value = 0.000000). **c-d,** Similar mutation patterns along captured regions were observed for Cohort 2 as for cohort #1 (Fig. 2e). **e**, To understand if observed variant frequencies are the result of clonal expansions or independent events, heterozygous variants were separated by allele. The logic behind this analysis is that if independently captured variants result from the same original event (i.e. a clone), then these variants should be found on the same allele. Alternatively, if variants result from independent events, then such variants should be frequently found on both alleles. By following linkage between variants and heterozygous SNPs, the two alleles can be distinguished. Shown here are the allele frequencies of variants found on either Allele 1 along the x-axis or Allele 2 along the y-axis (analyses are restricted to genomic segments from individuals containing heterozygous SNPs). As the variants adhere to a y=x line, they appear randomly distributed between both alleles, suggesting that variants detected represent multiple independent events rather than clonal expansions.

Extended Data Figure 4: Triplet prevalence in probed regions does not sufficiently explain base bias.

To understand how representative our total captured region was of the overall human genome, the trinucleotide sequence counts **a**, found within our 32 probes was compared to **b**, the overall trinucleotide counts found within hg19. CpG sites were less prevalently mutated in our samples than previously observed in other tissues and cancers. The lower incidence numbers of CpG mutations does not appear to be due to any effect of undersampling within our selected probe regions, as shown by **c**, the fold difference in the number of triplets found in our probed region and in the hg19 reference genome. Note that these analyses are of total sequence, not identified variants.

Extended Data Figure 5: 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 random 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 (**a-d**; p values ~1). 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. This suggests 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.

Extended Data Figure 6: Blood shows previously identified signatures but is different from 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.

Extended Data Figure 7: Oncogenic mutations do not show evidence of selection. As shown in Fig. 2f, 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.

Extended Data Figure 8: MMR^{MT} VAFs are elevated over parental frequencies. When compared to MMR sufficient HCT116 parental cell line genomic DNA, MMR deficient HCT116 cell DNA (R-Squared = 0.066023) contains substitution mutations at significantly elevated frequencies, as expected with DNA repair deficiencies (Fig. 3a-b).

Although most VAFs appear elevated within MMR deficient cells, the magnitude of increase was context dependent. Base substitutions altering **a-c**) C or G exhibited elevated allele frequencies in MMR^{MT} cells, but substantially less compared to **d-f**) T or A nucleotides, which exhibit much higher VAFs compared to parental.

Extended Data Figure 9: Base bias for cord blood individual #2 resembles MMR^{MT} Cells.

As for comparisons of MMR^{MT} and HCT116 parental cell lines, a cord blood donor showed a variant population that significantly deviated from expected VAFs (Fig. 3d). **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. Similar to the data in Extended Figure 8, base substitutions altering **b-d**) C or G nucleotides did not show elevated frequencies. As in the in the MMR^{MT} cells, **e-g**) T or A changes appear at elevated frequencies. Data from individual 19 looked similar to the data shown here, but is not shown. **h**, To ensure that the increased frequencies of variants are not the result of some experimental anomaly, the DNA from individuals #19 (not shown) and #2 was used in a second experiment. In the experimental repeat, the samples showed nearly identical mutational spectra, with similarly elevated levels of T or A changes. **i**, Indicative of experimental repeatability, when samples were freshly captured and sequenced using FERMI, the same individual was highly similar across experiments, and different individuals were less similar.