A comprehensive analysis of RNA sequences reveals macroscopic somatic clonal expansion across normal tissues

Supplementary Material

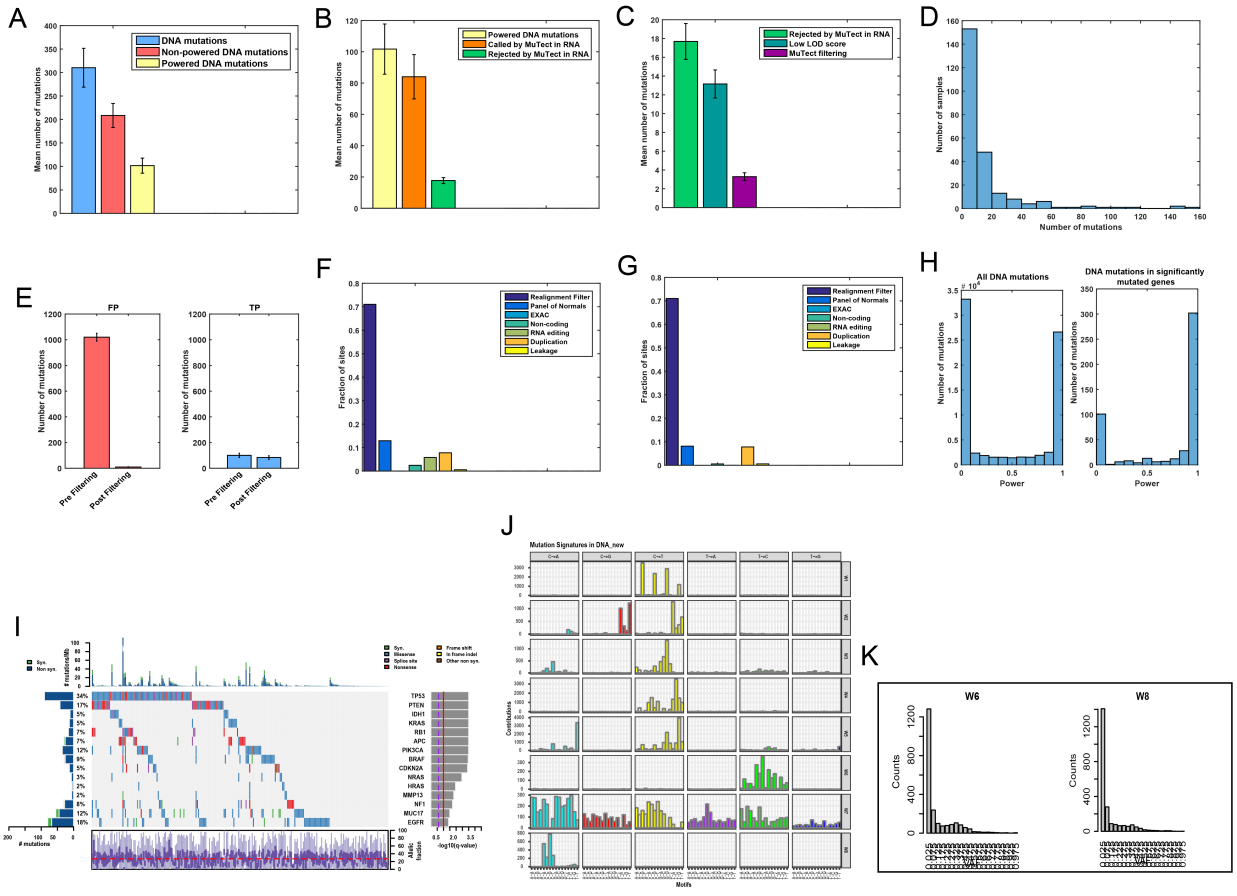

Figure S1: (A) Average number of DNA mutations in the studied cohorts is shown in blue. The average number of mutations for which we have sufficient power (≥95%) or non-sufficient power (<95%) for detection in the RNA is shown in red and yellow, respectively. Error bars represent the standard error of the mean. **(B)** Average number of DNA mutations with sufficient power to be detected in the RNA is shown in yellow. The average number of DNA mutations detected or not detected in the RNA by MuTect is shown in orange and green, respectively. Error bars represent the standard error of the mean. **(C)** Average number of DNA mutations rejected by MuTect when analyzing the RNA is shown in green. Out of this set, mutations filtered in the RNA due to a low LOD score are shown in dark green and those filtered by other MuTect filtering criteria are shown in purple. Error bars represent the standard error of the mean. **(D)** Distribution of the number of mutations per sample that are found in DNA, powered but not detected in the RNA. **(E)** Number of mutations detected before and after applying RNA-MuTect, for mutations that were not detected in the DNA (FP, left panel) and mutations that were detected in the DNA (TP, right panel). **(F)** Fraction of mutations filtered out by each filtering criteria. **(G)** Fraction of mutations that were uniquely filtered out by each filtering criteria. **(H)** Distribution of power to detect DNA mutations in RNA, for all mutations (left) and for mutations in significantly mutated genes (right). **(I)** CoMut plot derived from

a MutSigCV analysis based on mutations detected in DNA. Identified cancer genes (q-value < 0.05) with their frequency, mutation type and allele fraction in the studied cohort are presented. **(J)** Mutational signatures detected in the studied TCGA cohort based on DNA calls. The mutational signatures identified are: Aging (W1; COSMIC signature 1, cosine similarity = 0.94), APOBEC (W2; COSMIC signature 13, cosine similarity = 0.87), POLE (W3; COSMIC signature 10, cosine similarity = 0.88), UV (W4; COSMIC signature 7, cosine similarity = 0.99), Smoking (W6; COSMIC signature 4, cosine similarity = 0.86), MSI (W7; COSMIC signatures 15, cosine similarities of 0.92). W5 and W8 represent unannotated MSI and oxidative damage signatures. **(K)** Allele fraction distribution of mutations associated with signatures W5 and W8.

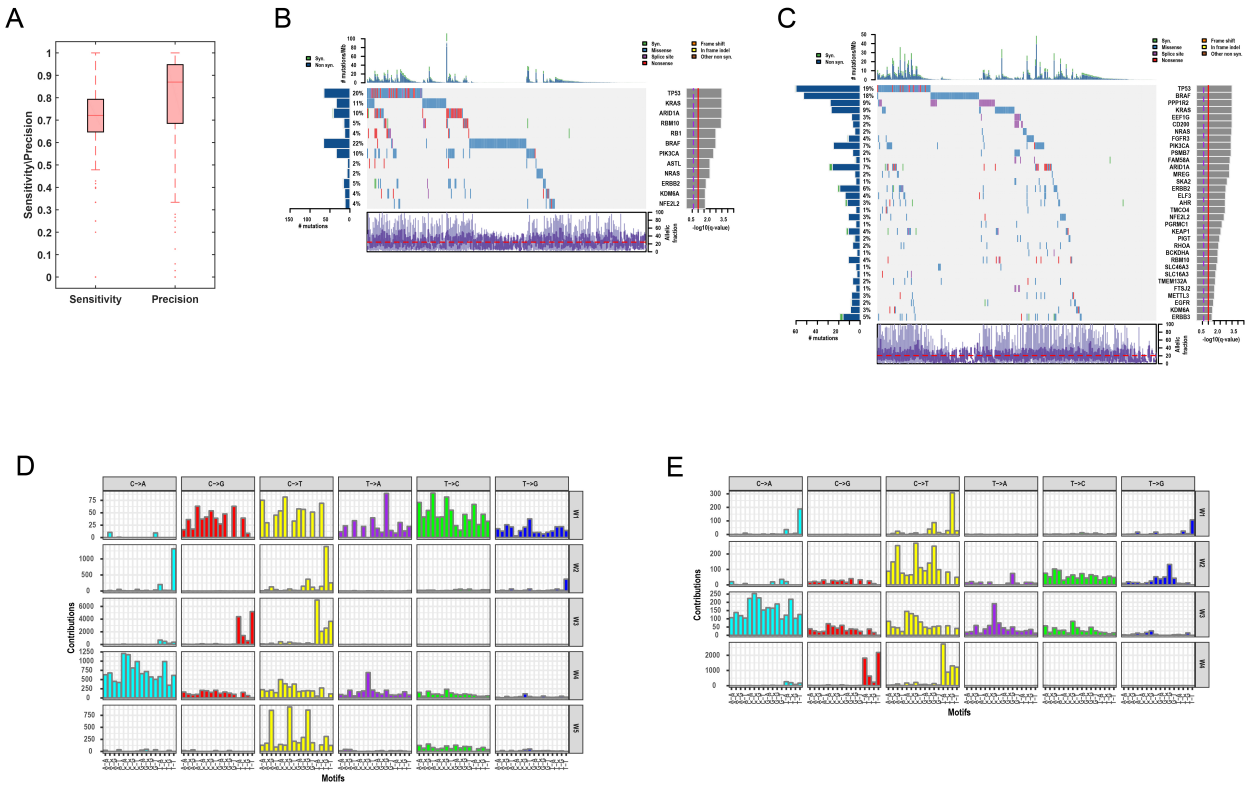

Figure S2: (A) Sensitivity and precision out of powered sites, across all studied cases in the validation cohort. Box plots show median, 25th and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the ' . ' symbol. **(B-C)** CoMut plot derived from a MutSigCV analysis based on mutations detected in the DNA (B) or RNA (C) of the validation cohort. Identified cancer genes (q-value < 0.05) with their frequency, mutation type and allele fraction in the studied cohort are presented. All genes except for RB1 and ASTL are identified as significantly mutated based on the RNA calls as well. **(D)** Mutational signatures detected in the validation TCGA cohort based on DNA calls. The mutational signatures identified are: POLE (W2; COSMIC signature 10, cosine similarity = 0.97), APOBEC (W3; COSMIC signature 2, cosine similarity = 0.84), Smoking (W4; COSMIC signature 4, cosine similarity = 0.96), Aging (W5; COSMIC signature 1, cosine similarity = 0.93). **(E)** Mutational signatures detected in the validation TCGA cohort based on RNA calls. The mutational signatures identified are: POLE (W1; COSMIC signature 10, cosine similarity = 0.9), a mixture of aging and DNA mismatch repair signatures (W3; COSMIC signature 1 and 6, cosine similarity = 0.78 and 0.8,

respectively), Smoking (W3; COSMIC signature 4, cosine similarity = 0.92), APOBEC (W4; COSMIC signature 2, cosine similarity = 0.8).

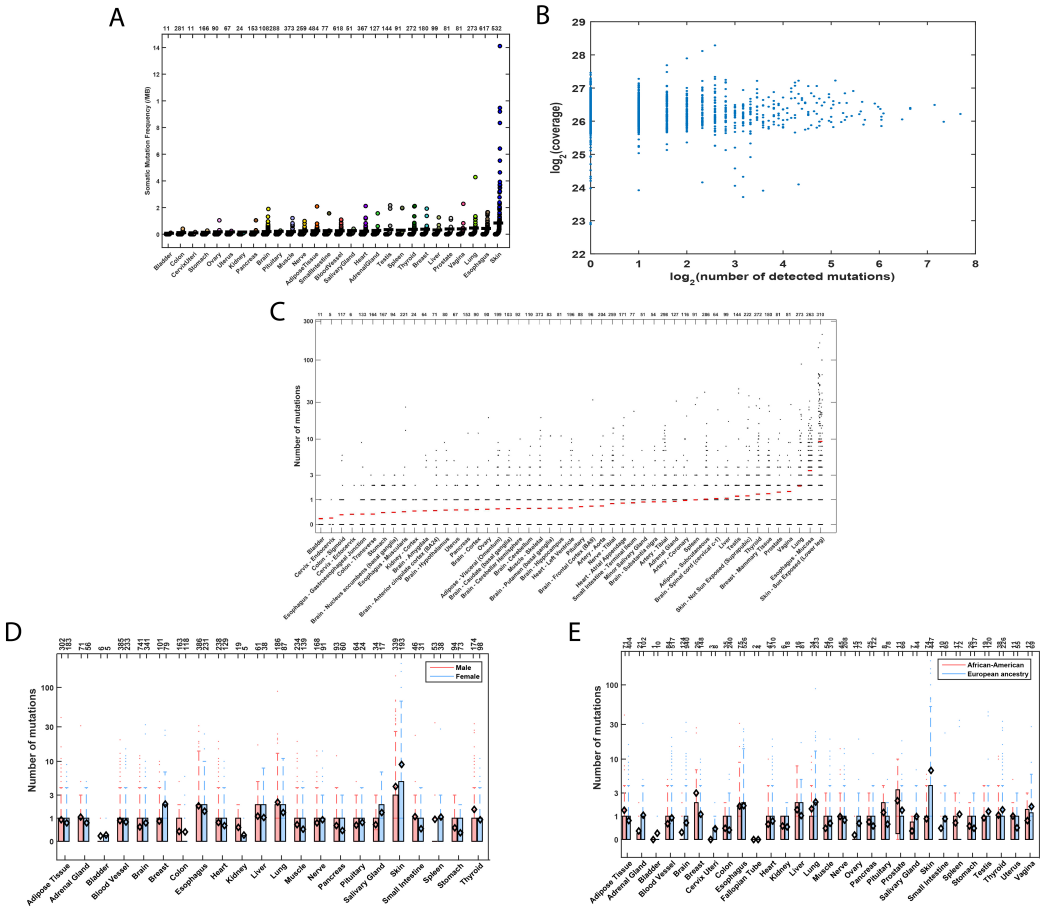

Figure S3: (A) Frequency of mutations detected in the RNA-seq of each tissue when separated by tissue type **(B)** $\log_2$(number of mutations detected in each sample+1) vs. the $\log_2$(coverage). Coverage equals to the total number of mapped reads. R=0.09, P-value=1.4×10^{-14} **(C)** Number of mutations detected in the RNA-seq of each tissue when separated by tissue site. The red horizontal line represents the mean over mutation number. The number of samples examined in each tissue is indicated at the top of the plot. **(D)** A boxplot describing differences in mutation number between males and females across different tissues. The black rhombus represent the mean of mutation number in each group. Number of samples in each group is indicated above. **(E)** A boxplot describing differences in mutation number between samples taken from individuals of European ancestry vs. those taken from individuals of African-American ancestry. The black rhombus represent the mean of mutation number in each group. Number of samples in each group is indicated above.

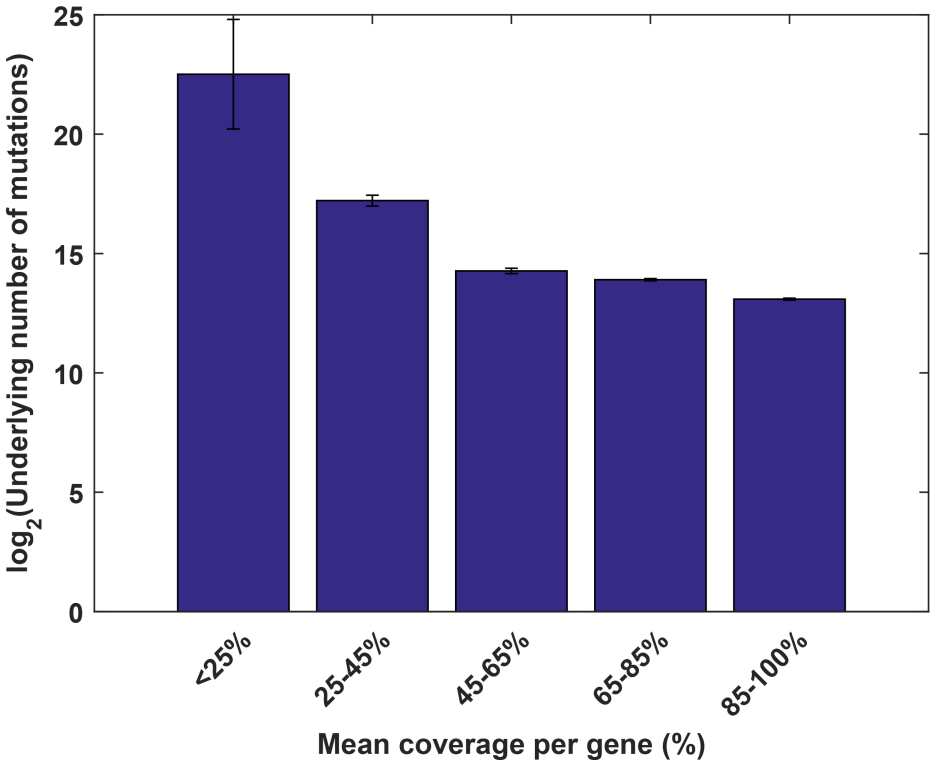

Figure S4: A bar plot showing the underlying number of mutations in log scale for different average coverage ranges per gene. A negative relation between the two factors is observed.

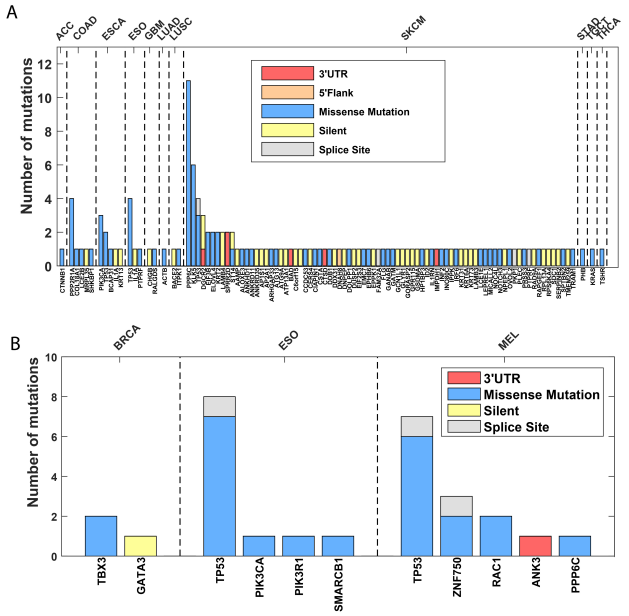

Figure S5: (A) Genes where the exact point mutation was found in a normal tissue and in the corresponding cancer tissue, based on MC3 data. The paired tissues reported are as follows: ACC -

Adrenocortical Carcinoma vs. Adrenal Gland; COAD – Colon Adenocarcinoma vs, Colon; ESCA – Esophageal Carcinoma vs. Esophagus; ESO – Esophageal Adenocarcinoma vs. Esophagus; GBM – Glioblastoma vs. Brain; LUAD – Lung Adenocarcinoma vs. Lung; LUSC – Lung Squamous Cell Carcinoma vs. Lung; SKCM – Skin Cutaneous Melanoma vs. Skin; STAD – Stomach Adenocarcinoma vs. Stomach; TGCT – Testicular Germ Cell Tumor vs. Testis; THCA – Thyroid Carcinoma vs. Thyroid. **(B)** Genes that were found to be significantly mutated in a given cancer type and mutated in the corresponding normal tissue. The paired tissues reported are as follows: BRCA – Breast Cancer vs. Breast; ESO – Esophageal Adenocarcinoma vs. Esophagus; MEL – Melanoma vs. Skin.

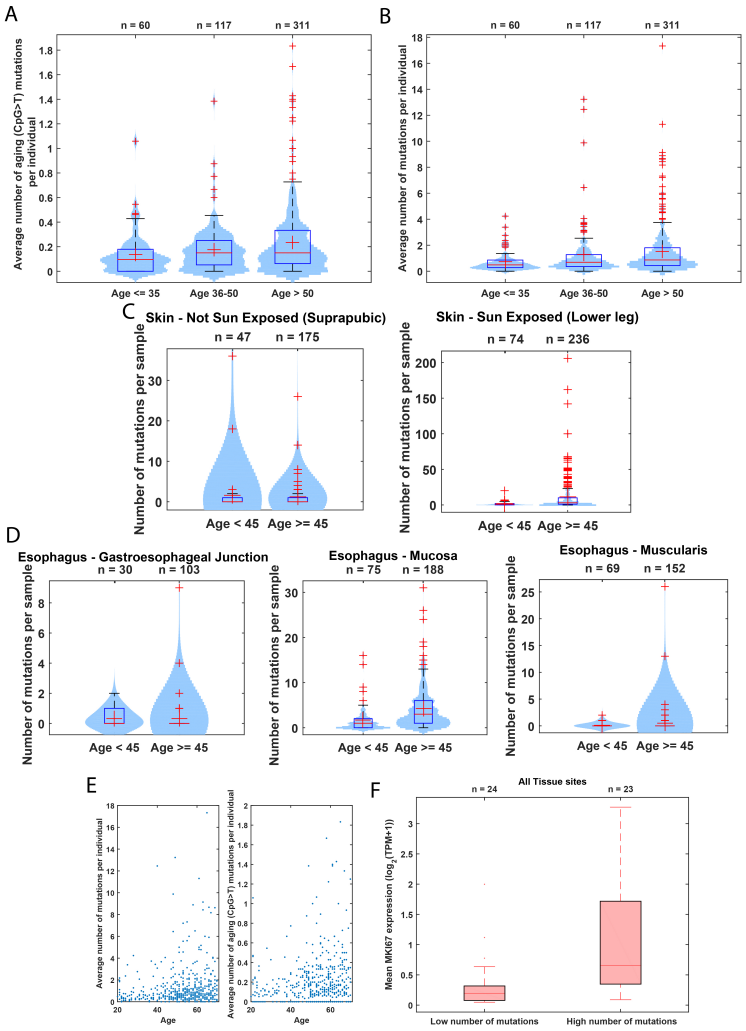

Figure S6: (A) Differences in the average number of aging-related mutations across different age groups. Violin plots show median, 25th and 75 percentiles of each group. Red crosses represent the outliers and black crosses represent the mean of each group. **(B)** Differences in the average number of total mutations across different age groups. Box plots show median, 25th and 75 percentiles of each group. Red crosses represent the outliers and black crosses represent the mean in each group. **(C)** A violin plot describing differences in mutation number between individuals younger or older than 45, for sun-exposed and non-exposed skin samples (one-sided Wilcoxon P-value = 3.9x10^{-8} and P-value = 0.04 for sun-exposed and non-

exposed). **(D)** A violin plot describing differences in mutation number between individuals younger or older than 45, for the different sub-regions of the esophagus: gastroesophageal Junction, mucosa and muscularis. The mucosa and muscularis showed a significant association with age (P-value = $8.4x10^{-9}$ and P-value = 0.02). **(E)** Scatter plot describing the average number of mutations (left panel) and average number of aging mutation per individual (right panel), vs. the age of the individual **(F)** Difference in the expression of *MKI67* in all tissue sites (P-value = $1.2*10^{-4}$). Tissues are grouped to those with lower or higher total mutation number in respect to the median. We considered only tissues with more than 10 samples (47 out of 50).

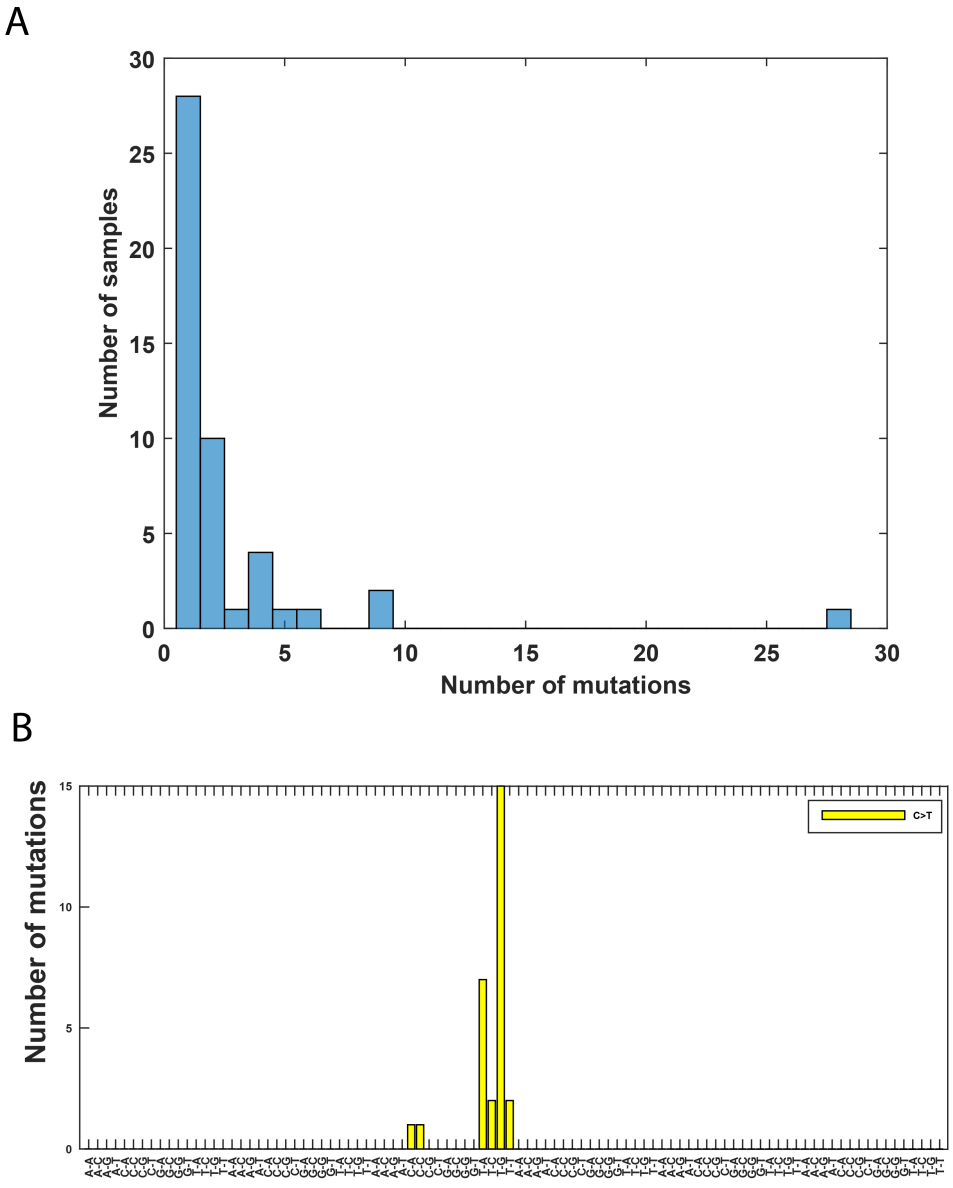

Figure S7: (A) Distribution of mutation number in 48 samples taken from the vagina tissue. **(B)** Mutation spectrum for the vagina sample with a relatively high mutation number. Only C>T mutations with the tri-nucleotide context described in the figure were detected.

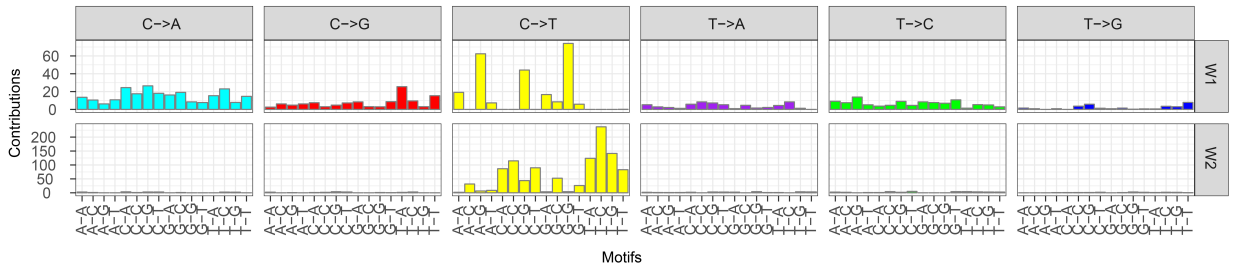

Figure S8: Mutational signatures identified for TCGA simulated data, where the distribution of mutation number per sample is identical to that found in GTEx. Only two signatures, the UV signature (W2), and an additional one most closely related to the aging signature (W1) but with a relatively low cosine similarity.

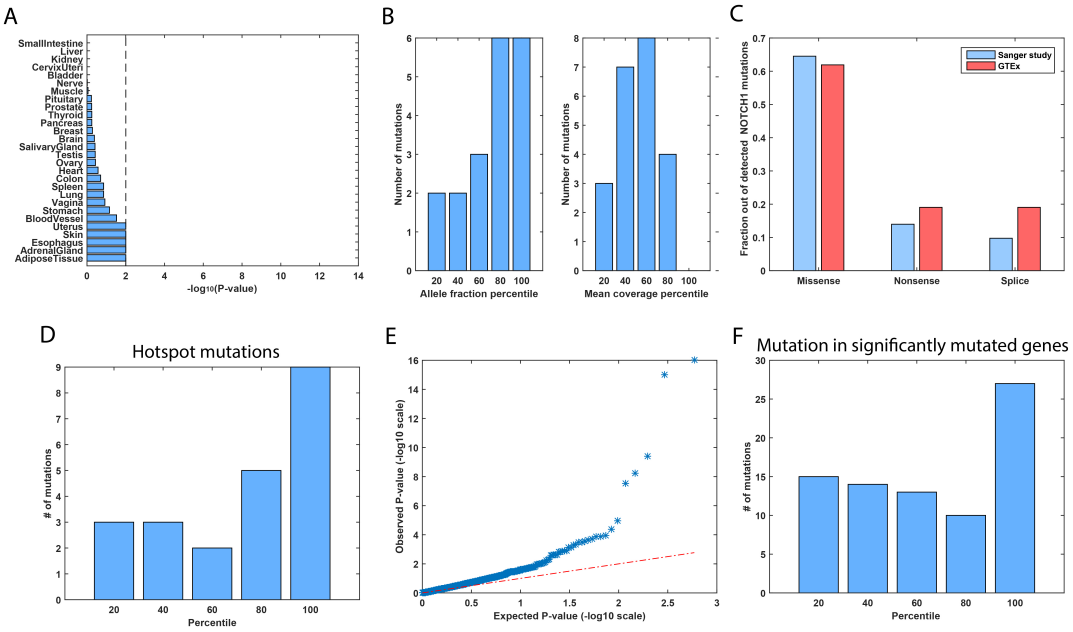

Figure S9: (A) -$\log_{10}$(empirical P-value), testing for enrichment for non-silent mutations in CGC genes in each tissue (**Methods**). The dashed line represents the significance threshold based on FDR with $\alpha = 0.1$. **(B)** left panel: Number of *TP53* mutations that their allele fraction falls above the indicated percentile, relative to all other mutations in the samples they were detected in; right panel: Number of *TP53* mutations that their average coverage per base falls above the indicated percentile, relative to all other mutations in the samples they were detected in. **(C)** Fraction of different *NOTCH1* mutation types out of all *NOTCH1* detected mutations, in the Sanger (Martincorena et al., 2015) study and in the GTEx data. **(D)** Number of hotspot mutations that their allele fraction falls above the indicated percentile, relative to all other mutations in the samples they were detected in. **(E)** QQ-plot for pan-normal MutSigCV analysis based on P_{CV} values. **(F)** Number of mutations in significantly mutated genes that their allele fraction falls above the indicated percentile, relative to all other mutations in the samples they were detected in.

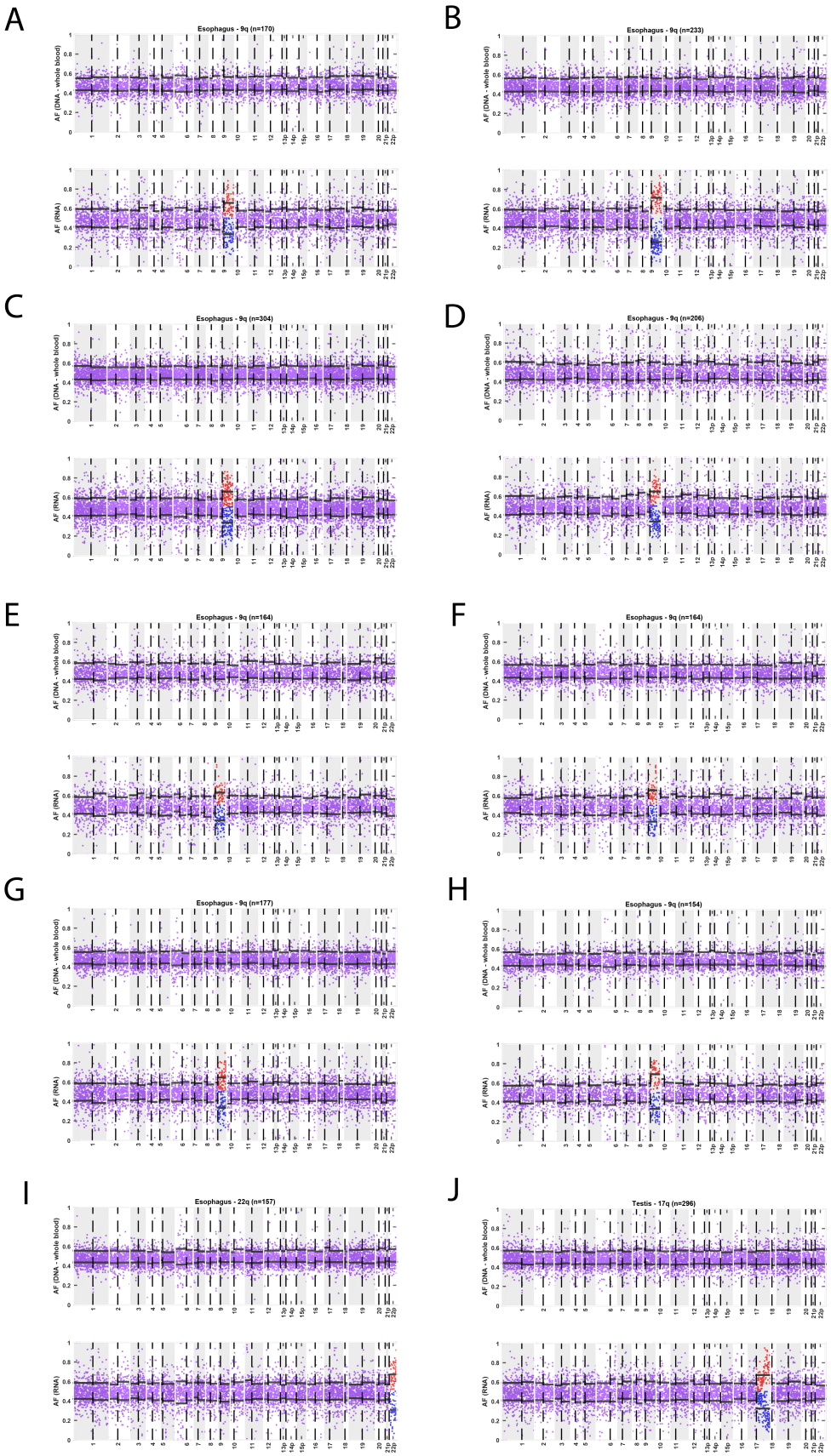
A
Esophagus - 9q (n=170)
B
Esophagus - 9q (n=233)
C
Esophagus - 9q (n=304)
D
Esophagus - 9q (n=206)
E
Esophagus - 9q (n=164)
F
Esophagus - 9q (n=164)
G
Esophagus - 9q (n=177)
H
Esophagus - 9q (n=154)
I
Esophagus - 22q (n=157)
J
Testis - 17q (n=296)
AF (DNA - whole blood)
AF (RNA)

Figure S10: (A-H) Allelic imbalance in chrmosme 9q of 8 normal esophagus samples. The top panel in each sub-figure shows the allele fraction of heterozygous sites based on DNA from a matched-blood sample. The bottom panel shows the allele fraction of heterozygous sites based on RNA from the same sample. The black horizonal lines indicate the mean allele fraction per chromosome arm of sites with allele fraction smaller or greater than 0.5. **(I)** Allelic imbalance in chrmosme 22q of a normal esophagus sample. **(J)** Allelic imbalance in chrmosme 17q of a normal testis sample.

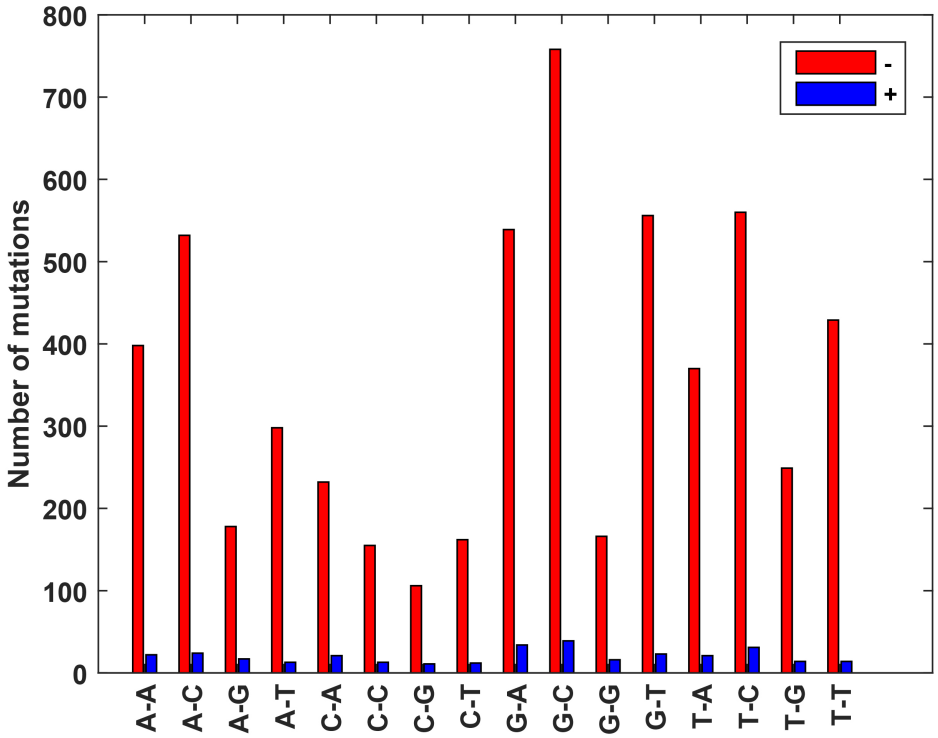

Figure S11: Number of C>A / G>T mutations on transcribed (blue) and non-transcribed (red) strands in all possible trinucleotide contexts. A strand bias towards the non-transcribed strand is observed (Binomial P-value < 4.9×10^{-324}).