Motif-Centered Analyses Elucidate Mutagenic Mechanisms Operating in Normal Human Tissues

Safia Mahabub Sauty[1], Yun-Chung Hsiao[1], Leszek J. Klimczak[2], Dmitry A. Gordenin[1#]

[1]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, Durham, NC 27709; [2]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, Durham, NC 27709

#Corresponding author: gordenin@niehs.nih.gov.

Running title: Detection of mutational motifs in normal tissues

Abstract

Somatic mutations are inevitable in human genomes and can lead to tumorigenesis. Large-scale bioinformatic analyses of mutation catalogues have unveiled several mutational processes operating in cancer genomes, however, our understanding of baseline mutagenesis in non-cancerous normal cells is scarce. Here, we analyzed the mutation profiles of 11,949 normal samples across 25 tissue types obtained from whole-genome and whole-exome sequencing datasets. We applied stringent statistical hypothesis for detecting enrichment and enrichment-adjusted Minimal Estimate of Mutation Load (MEML) in trinucleotide motifs preferred by known mutagenic processes. We found that several mutational motifs previously identified in human cancers are readily detected in cancer-free tissues. Samples enriched with C to T mutations in nCg motif associated with clock-like spontaneous deamination of meCpG were detected across all tissues. We revealed another ubiquitous clock-like motif in several tissue types - T to C substitutions in aTn motif previously associated with exposure to small epoxides and other S_N2 electrophiles. Analyses of UV-associated yCn motif and its sub-motifs, as well as a minor UV-motif nTt revealed UV-mutagenesis only in skin. APOBEC-induced C to T and C to G mutations in tCw motif were enriched in bladder, lung, small intestine, liver, and breast with preference for

APOBEC3A-like mutagenesis in most. Non-cancerous diseased samples showed significantly higher, age-independent accumulation of the clock-like motifs aTn and nCg compared to healthy tissues. Together, our analyses elucidated several ongoing mutagenic processes in normal human tissues and provided a robust analytical framework for identifying mutagenic sources from low-burden mutation catalogues.

Introduction

Mutations are inevitable in genomes of all species, including humans. Genome sequencing breakthrough more than a decade ago brought realization of massive burden of *de novo* mutations in human cancers (Alexandrov et al. 2013, Lawrence et al. 2013, Roberts and Gordenin 2014). Earlier research also led to propose that mutation accumulation in non-cancerous normal somatic cells can be carcinogenic (Ames et al. 1973, Loeb et al. 1974). Somatic mutations can also result in disease other than cancer (Erickson 2003, 2014, 2016). The possible roles of somatic mutagenesis in cancer and other diseases led to questions about the rates and causes of somatic mosaicism and genome instability in humans. Mutation catalogs of somatic cells and tissues can be obtained from conventional whole genome sequencing of cell clones, deep sequencing of micro biopsies, sequencing of multiple single cells from a tissue or by accurate single-molecule sequencing from a non-clonal population of cells derived from a tissue ((Martincorena and Campbell 2015, Saini and Gordenin 2018, Evrony et al. 2021, Maslov and Vijg 2023, Zhang et al. 2024) and references therein). Catalogs of somatic mutations in normal (non-cancerous) human tissues generated in dozens of independent studies were accumulated in the SomaMutDB database (https://vijglab.einsteinmed.edu/SomaMutDB) (Sun et al. 2022). The collected large volume of mutation data enabled mutational signature extraction to be applied to the entire SomaMutDB dataset, to its' sub-sets including samples from specific tissues, to other datasets and then to use these signatures for addressing specific hypotheses (see e.g., (Moore et al. 2021, Yaacov et al. 2023)). Agnostic signature extraction turned to be a powerful and popular

tool allowing indication of the presence of known as well as yet unknown mutational processes operating in human cancers (Alexandrov et al. 2013, Alexandrov et al. 2020). However, at the very inception of this method it had been stated that "*Parameters* [of signature deciphering] *to which solutions are sensitive include the number of operative mutational processes, the strength of their exposures, the degree of difference between mutational signatures, the number of analyzed cancer genomes, the number of mutations per cancer genome, and the number of mutation types that are incorporated into the model*" (Alexandrov et al. 2013). These methods' limitations can be significant in application to signature extraction from the mutational catalogs of normal somatic cells and tissues, where the number of mutations per sample and the sample size of a cohort can be small. Standard approach with agnostic signatures derived from catalogs of somatic mutations in normal tissues is to align somatic signatures with canonical COSMIC signatures (https://cancer.sanger.ac.uk/signatures/sbs/ and (Alexandrov et al. 2020)) using signature extractor tools and/or calculating cosine similarity (Moore et al. 2021, Yaacov et al. 2023, Zhang et al. 2024). This approach allows to highlight presence of some known mutagenic processes but does not assign P-values to the prevalence of signature(s) in an individual sample.

We previously developed the knowledge-based motif-centered analytical pipeline P-MACD (https://github.com/NIEHS/P-MACD) complimentary to agnostic signature extraction. The pipeline is set to calculate enrichment in mutation catalog of an individual sample with mutations in trinucleotide motif(s) that have been identified by mechanistic research to be preferred by a single or by a group of related mutagenic factors (see Materials and Methods). The parameters of motif-specific mutagenesis in a sample are just the two odds-ratios, which allows to calculate sample-specific statistics, apply FDR correction, and estimate mutation load caused by a motif-preferring mechanism in each sample of a cohort. Our and others' prior studies have demonstrated several utilities of motif-centered analyses. For example, it allowed to distinguish between APOBEC3A and APOBEC3B mutation load in human cancers (Chan et al. 2015, Petljak

et al. 2019, Petljak et al. 2022), detect APOBEC mutagenesis in a subset of lung tumor samples from smokers where agnostic extraction did not indicate APOBEC mutagenesis (Zhang et al. 2025), identify minor UV-induced base substitution and InDel mutational motifs as corroborating evidence for UV-mutagenesis in human skin (Saini et al. 2021). Motif-centered analyses have revealed statistically significant presence of a motif identical to COSMIC signature SBS16 in all types of human cancers, while the agnostic extraction assigned SBS16 to only two (at most three) cancer types (Hudson et al. 2023). Trinucleotide motifs derived from mechanistic research of several mutagenic factors are summarized in Table 1.

Here we report findings revealed by motif-centered pipeline applied to mutation catalogs of over ten thousand whole-genome and whole-exome sequenced (WGS and WES) samples from a wide range of normal human tissues obtained from healthy and disease-carrying individuals. We found that enrichment with mechanism-associated motifs previously detected in human cancers also exists in normal tissues. We also revealed association of several motifs with tissue type, donor age, and disease status. The analytical framework presented in our study can be applied to any mutation catalogs of individual somatic samples as well to large sample cohorts.

Results

Study design

We analyzed the mutation calls from WGS and WES samples of non-cancerous (i.e., normal) human tissues. The mutation calls were retrieved from the somatic mutation database SomaMutDB (https://vijglab.einsteinmed.edu/SomaMutDB) (Sun et al. 2022), and from the supplementary materials of several papers not included into SomaMutDB (see Materials and Methods) (Supplemental Table S1a). The dataset of mutation calls was categorized into two groups, healthy and diseased normal tissues. The mutation calls in the healthy cohort are from 10,625 samples of 777 individual donors, spread across 25 organs (Supplemental Table S1b).

8,054 of these samples are WGS with total 7,623,352 mutations (sample mutation loads range from zero to 34,629). 2,571 of healthy normal samples are WES with total 179,208 mutations (sample mutation loads range from zero to 5,957). Detailed information of both WGS and WES healthy samples, including their mutation loads, are presented in Supplemental Table S1b. The mutation calls in the diseased cohort are from 1,323 samples of 223 Individual donors, spread across 6 tissue types and 15 different disease conditions (Supplemental Table S1c). 1,226 of the diseased samples are WGS (with 3,243,116 total mutations; sample mutation loads range from zero to 18,243) and 96 samples are WES (with total 65,843 mutations; sample mutation loads range from 9 to 2,931). Detailed information of the diseased cohort is available in Supplemental Table S1c.

We performed motif-centered analyses to find enrichment of known mutagenic processes and quantify their presence in individual samples. To conduct motif-centered analyses of the mutation calls, we first curated a list of trinucleotide motifs preferred by known mutagenic mechanisms and supported by experimental evidence and mechanistic knowledge (Table 1). In order to resolve potential overlap between motifs centered around the same base substitution, we built a non-overlapping sub-motif, where needed (Supplemental Table S2).

Table 1: List of knowledge-derived mutational motifs.

Motif name	Reference trinucleotide	Mutant trinucleotide	Motif for SBS correlation	Mechanism	References
aTn	aTn	aCn	aTn→aCn	Epoxide	(Hudson et al. 2023)
yCn	yCn	yTn	yCn→yTn	UV	(You et al. 2001, Saini et al. 2016, Saini et al. 2021)
nCg	nCg	nTg	nCg→nTg	Spontaneous deamination of meCpG	(Pfeifer 2006, Saini et al. 2016, Saini et al. 2021)
nTt	nTt	nCt	nTt→nCt	UV (minor)	(Saini et al. 2016, Saini et al. 2021)
tCw	tCw	tTw	tCw→tTw	APOBEC	(Roberts et al. 2013)
tCw	tCw	tGw	tCw→tGw	APOBEC	(Roberts et al. 2013)
tgC	tgC	tgA	tgC→tgA	Redox	(Degtyareva et al. 2023))
hTg	hTg	hGg	hTg→hGg	SN_2 alkylation	(Saini et al. 2020)
aCy	aTy	aCy	aCy→aTy	SN_1 alkylation	(Saini et al. 2020)
gCn	gCn	gAn	gCn→gAn	Acetaldehyde	(Vijayraghavan et al. 2022)
cTg	cTg	cAg	cTg→cAg	Aristolactam (from Aristolochic acid)	(Rosenquist and Grollman 2016)

We evaluated statistically significant over-representation of 11 knowledge-based trinucleotide motifs using motif-centered analysis tool P-MACD (https://github.com/NIEHS/P-MACD). P-MACD calculates enrichment of mutations in the trinucleotide context relative to random mutagenesis. It assigns sample-specific P-values and calculates enrichment-adjusted Minimum Estimate of Mutation Load (hereafter referred as MEML) for samples that have a statistically significant enrichment of mutational motif after correction for multiple-hypothesis testing. By forming a stringent statistical hypothesis and performing sample-specific analyses, we extracted the MEML that can be attributed to a specific mutagenic mechanism. Conveniently, non-zero MEML value indicates statistically significant enrichment with a motif, so these two terms can be used interchangeably.

We also compared the mutation profiles of different normal tissues with the published COSMIC reference SBS signatures derived from large cohorts of cancer mutation catalogs.

Mutation spectrum and the contribution of COSMIC SBS signatures in normal tissues

Dataset of mutations in healthy normal tissues contained a total of 7.8 million single base substitutions (SBS) identified by whole exome or by whole-genome sequencing (hereafter referred as WES or WGS, respectively). Median SBS counts in WGS healthy tissues ranged from 22 to 4,025, while counts in WES healthy tissues ranged from 9 to 499 (Fig. 1A, Supplemental Table S3a). Similar to prior reports of genome-wide mutagenesis in multiple cancer types (Alexandrov et al. 2013, Lawrence et al. 2013, Alexandrov et al. 2020) we observed highest fraction of C to T substitutions across all organs. Mutation catalogs of normal tissues with noncancer diseases contained 3.3 million substitutions with median SBS count ranging 29 to 6,026 in WGS samples and 35 o 1,135 in WES samples (Fig. 1B, Supplemental Table S3b).

We listed the total mutation calls from each tissue and disease types as 96-trinucleodtide motif profiles using SigProfilerMatrixGenerator (Khandekar et al. 2023), and fitted them with already published COSMIC SBS signatures derived from PCAWG cancer catalogue (Alexandrov et al. 2020) to estimate the relative contribution of cancer-associated signatures in the mutation loads of normal tissues (Supplemental table S3c, S3d). We detected the contribution of SBS1 associated with meC deamination in meCpG across most tissues and disease types (Supplemental Fig. S1, S2), however relative contribution of SBS1 was small. The signatures with the largest relative contributions detected across most tissue and disease types were SBS24 (proposed to associate with aflatoxin exposure (Chawanthayatham et al. 2017, Alexandrov et al. 2020), found only in liver cancers), SBS32 (proposed to associate with azathioprine exposure (Inman et al. 2018, Alexandrov et al. 2020), found only in 3 samples of different cancer types), and SBS39 (no proposed etiology, found in 0.3-7% samples of seven cancer types). We further calculated the cosine similarity between published SBS signatures and the mutation profiles of normal tissues. The cosine similarity values ranged from zero to 0.74 with a median cosine value of 0.249 in the healthy cohort (Supplemental Table S3e). In the diseased cohort the cosine similarities ranged from 0 to 0.723 with a median value of 0.248 (Supplemental Table S3f). We

did not detect any strong similarity (usually postulated for cosine similarity value ≥0.8) with any of the published signatures. SBS3, SBS5, SBS39, SBS40, and SBS89 were found to have the strongest and most prevalent cosine similarity across most tissue and disease types (Fig. 1C,D). Altogether, assessment of the prevalence of cancer SBS signatures in totaled somatic mutation catalogs of normal tissue types did not reveal strong presence of any signature with known etiology except SBS1. As outlined in the Introduction, we proposed that sample-specific motif-centered evaluation may reveal and quantify contributions of known mutagenic sources and mechanisms on somatic mutagenesis in humans where signature approach may still require more data.

Mutational motifs derived from mechanistic research are enriched in various human tissues

We used P-MACD pipeline to assess enrichment with eleven selected trinucleotide mutational motifs. These mutational motifs were previously associated with mutagenic mechanisms indicated in mechanistic studies and were also found in human cancers and in normal human skin (Table 1). Strongest prior evidence supported nCg motif characteristic of cytidine deamination in meCpG dinucleotides (Pfeifer 2006), tCw motifs of APOBEC mutagenesis (Roberts et al. 2012, Roberts et al. 2013, Mertz et al. 2022), aTn motif of exposure to small epoxides (Hudson 2023), and yCn motif of UV-mutagenesis (You et al. 2001, Saini et al. 2016, Saini et al. 2021). We also extended this analyses to the WES samples that have 100-fold less mutations compared to the WGS samples (Supplemental Fig. S3A, B).

Analysis of mutational motif enrichment indicated wide-spread presence of the two motifs associated with mutagenic processes in different types of morphologically normal healthy tissues (Fig. 2A). The aTn and nCg, were detected in 4%-100% and 43%-100% of the WGS samples, respectively (Supplemental Table S4a). Statistically significant enrichment with nCg was also detected in 9%-93% of WES samples, while 0%-20% WES samples showed aTn MEML, further

validating the robustness of the motif-centered analysis (Supplemental Table S4a, Supplemental Fig. S3C). aTn and nCg were also detected across all diseased tissues (Fig. 2B). aTn was 3-100% of the WGS samples and 4%-12% of the WES samples. nCg MEML was detected in 26%-100% of WGS and 94%-100% of WES diseased samples (Fig. 2B, Supplemental Fig. S3D, Supplemental Table S3d). Other motifs were detected only in some organs and diseased tissues, suggesting tissue-specificity of the underlying mutagenic processes.

Below we dissect the features of individual motifs and their baseline prevalence in normal healthy tissues. We then explore association of motif-specific mutagenesis with disease state of noncancer tissues.

The small epoxide mutagenesis motif aTn→aCn is enriched in samples across all human tissues and accumulates along the lifetime.

The aTn→aCn motif (referred to aTn throughout this paper) was experimentally identified as a defining feature of hypermutation caused in a yeast sub-telomeric single-stranded DNA (ssDNA) model system in response to treatment with small epoxide glycidamide – an electrophile reacting with nucleobases via S_N2 mechanism (Hudson et al. 2023). Glycidamide-induced hypermutation in ssDNA showed clear overrepresentation of A→G substitution in nAt context (T→C in aTn in reverse complement). The base preference of glycidamide showed similarity with the chemical preference of S_N2 reacting electrophiles in ssDNA and with the mutagenic preference of another electrophilic alkylating agent reacting via S_N2 mechanism, methyl methanesulfonate (MMS), supporting the hypothesis about an aTn motif preference not only by small epoxides, but by a broader class of S_N2 reacting electrophiles ((Saini et al. 2020, Hudson et al. 2023) and references therein). Analysis of mutation catalogs from PanCancer Analysis of Whole Genomes (PCAWG) (Consortium 2020) detected widespread enrichment and clock-like accumulation of aTn motif in multiple cancer types. Ubiquitous nature of aTn motif further suggested that it may be a feature of a broader class of mutagens instead of glycidamide alone.

Prevalence of aTn motif was shown to have strong positive correlations with mutation loads assigned to COSMIC signatures SBS1, SBS4, SBS5 and SBS16 in tumors where these signatures were detected. In cancers with smoking as a risk factor, aTn mutation load showed a strong and significant association with the smoking history of the patients. Smoking-related aTn mutation load in lung tumors largely exceeded age-dependent presumably endogenous level of mutagenesis. Age-dependent aTn accumulation was also detected in healthy skin and brain cells (Hudson et al. 2023).

We detected aTn motif prevalence in morphologically healthy samples across all tissue types. While aTn→aCn MEML contributed only a modest fraction to a median 9.03% (95% confidence interval: 6.7%-10.4%) of total AT pair mutations (A→C, A→G, A→T, including reverse complements), its' statistically significant enrichment was detected in WGS samples across all tissue types (Fig. 3A, Supplemental Table S5a). This indicates the motif's utility for evaluating the level of epoxide- or even broader class of S_N2 electrophile-induced mutagenic exposure. The aTn motif-associated mutagenesis was also detected in WES samples, albeit at lower prevalence as compared to WGS mutation catalogs. Multiple samples showing non-zero aTn MEML were detected in all tissue types except epidermis and heart, (Fig. 3B, Supplemental Table S5b).

T→C substitutions are central to aTn motif, however other mutagenic processes causing the same substitutions can confound statistical evaluation of aTn enrichment. Another motif centered around T→C, nTt→nCt (nTt in abbreviated format) is generated by the error-prone trans-lesion synthesis (TLS) across UV-induced cyclo-pyrimidine dimers (CPDs) formed by two adjacent thymines. It was revealed as a secondary minor UV-associated motif (You et al. 2001, Saini et al. 2016, Saini et al. 2021). Importantly, aTn motif is a part of nTt motif and vice a versa – i.e. the motifs overlap (See Supplemental Table S2). Such an overlap would not allow an unambiguous assignment of mutagenic mechanism, UV or S_N2 reacting electrophiles, because of mutually confounding effects of respective diagnostic enrichment and MEML estimates of aTn

and nTt motifs. This difficulty can be resolved using non-overlapping parts (i.e., sub-motifs) of mutually confounding motifs (Supplemental Table S2). To identify the presence of bona fide aTn contribution into MEML, we performed Spearman correlation tests between MEML counts of the aTn and its sub-motif, aTr (Supplemental Table S5c). In the aTr motif, thymine is followed by either an adenine or a guanine which eliminates any contribution of UV-mediated thymine-thymine dimers of the nTt motif. Significant positive correlation was detected between aTn and aTr MEML across all tissue types except appendix, epidermis, and tonsil (Fig. 3C, Supplemental Fig. S4A). Lack of significant aTn-aTr correlation in epidermis and tonsil can be attributed to the low number of samples with both aTn and aTr MEML (Supplemental Fig. S4B), however confounding effect of other mutagenic mechanisms cannot be excluded, especially in epidermis where UV mutagenesis could occur. Correlation analyses excluding either aTn MEML=0 or aTr MEML=0 further confirms that the strong contribution of aTn MEML in most tissues cannot be explained by confounding UV-mutagenesis (Supplemental Fig. S4B, Supplemental Table S5c).

Next, we asked if the aTn MEML accumulation shows clock-like feature across normal human tissues as previously identified in cancers and in healthy skin (Hudson et al. 2023). We calculated mean MEML load in each donor and performed spearman correlation analyses with donor age within each tissue (Supplemental Table S5d). We detected significant positive correlation with age in all tissues where aTn and aTr was detected, except in bladder (Fig. 3D, Supplemental Table S5e). Of note, aTr shows a stronger correlation with age than aTn in dermis, suggesting a UV-like component (nTt) that is not age dependent. Correlation analyses between donor age and aTn or aTr MEML excluding zero MEML values reproduced same finding (Supplemental Fig. S4C). Together, our analyses showed similar features of aTn motif in healthy tissues as detected in tumors and generalizes its' cock-like feature to include multiple tissue types.

Clock-like mutagenesis caused by deamination of 5-methylcytosine motif nCg→nTg is highly enriched in multiple human tissues

In mammalian genomes, 70%-80% of the cytosines in CpG dinucleotides (nCg motifs) are methylated, and spontaneous 5-methylcytosine deamination results in C→T substitutions (Pfeifer 2006). This nCg motif mutagenesis is ubiquitously detected across all tumor types in WES and WGS TCGA the PCAWG catalogs, shows clock-like accumulation, and corresponds with COSMIC signature SBS1 (Alexandrov et al. 2013, Alexandrov et al. 2020).

In our cohort of WGS samples in healthy tissues, we detected nCg MEML across all tissue types with 14.15% median nCg→nTg MEML contribution in total CG-pair mutations (95% confidence interval: 9.6%-20.7%) (Fig. 4A). WES samples showed 15.5% median nCg contribution in CG-pair mutations (95% confidence interval: 3.7%-40.9%) (Fig. 4A) (Supplemental Table S6a). Samples with nCg MEML were detected across all tissue types with sample-specific MEML ranging from zero to 2,010 in WGS samples, and from zero to 1,456 in WES samples (Fig 4B, Supplemental Table S6b). Median nCg MEML detected across WGS samples is 61 (95% confidence interval: 59-63) and that in WES samples is 3 (95% confidence interval: 3-4).

Multiple mutagenic processes cause cytosine deamination in mammalian tissues including CpG deamination, repair of UV-mediated cyclobutane pyrimidine dimers (CPDs), and activities of cytidine deaminases like AID and APOBEC (Roberts and Gordenin 2014). The trinucleotide base composition in nCg motif overlaps with base preference by UV mutagenesis, which forms pyrimidine-pyrimidine dimers in yCn context. To separate the contribution of yCn and detect true nCg MEML in tissues, we used the approach of sub-motif analyses and split the nCg motif into two sub-motifs, rCg and yCg (Supplemental Table S2). The mutable cytosine in rCg is preceded and followed by a purine, therefore excluding the possibility of CPD formation by UV-damage. yCg on the other hand has a pyrimidine preceding the mutable cytosine and can be converted to a thymine by both spontaneous and UV-mediated deamination. Performing Spearman correlation analyses of nCg MEML with both rCg MEML and yCg MEML allowed us to identify tissues that do not show true spontaneous CpG deamination (Supplemental Table S6c).

We detected strong positive nCg-rCg and nCg-yCg correlations in all tissues and sequencing groups, except in dermis and epidermis where nCg-rCg correlation is absent (in epidermis) or weaker (in dermis) compared to nCg-yCg correlation (Fig. 4C, Supplemental Fig. S5A, S6A).

Correlation analyses excluding samples that had either motif MEML=0 showed strong nCg-rCg and nCg-yCg correlations in both epidermis and dermis samples (Supplemental Fig. S5B). To further dissect the effect of UV in skin samples, we analyzed the yCh→yTh motif which overlaps with UV-preferred yCn motif but excludes CpG motifs (Supplemental Fig. S5C, Supplemental Tables S6b, S6c). We found a very strong nCg-yCh correlation (WGS: Spearman's ρ=0.9, q-value=<0.0001; WES: Spearman's ρ=0.85, q-value=<0.0001) in epidermis samples. Dermis samples showed a moderately strong nCg-yCh correlation (WGS: Spearman's ρ=0.6, q-value=<0.0001). Based on these sub-motif analyses we concluded that the spontaneous deamination of nCg is detectable but can be heavily confounded by UV-mutagenesis in skin samples.

Next, we investigated the clock-like feature of meCpG deamination as previously detected in cancers (Alexandrov et al. 2013). For donors with known age, we calculated average donor-specific MEML value for nCg motif and its sub-motifs, yCg and rCg (Supplemental Table S6d). Spearman correlation analyses showed significant positive correlation between donor age and all three motifs (nCg→nTg, rCg→rTg, yCg→yTg) in tissues where true spontaneous meCpG deamination activity was indicated by sub-motif analyses described above (Fig. 4D, Supplemental Fig. S6B, Supplemental Table S6e).

Mutation motifs yCn→yTn and nTt→nCt detected in several tissues can support UV mutagenesis only in skin

UV radiation is a common and potent environmental mutagen that facilitates many types of genome changes including SBSes. UV causes covalent bond formation between two adjacent

pyrimidines that results in cyclobutane pyrimidine dimers (CPD) and pyrimidine 6–4 pyrimidone (6-4PP) formation, with CPDs contributing to >80% of the UVB-mediated mutations in mammalian cells (You et al. 2001). C→T mutations in the yCn context are the most frequent result of UV lesions. The yCn→yTn mutations can result from direct error-prone copying of a cytosine within a CPD by TLS polymerases. Importantly, spontaneous deamination of cytosine changing it to uracil is increased several orders of magnitude in CPDs (Peng and Shaw 1996) and can result in C→T substitutions after accurate copying by TLS polymerases and another round of replication. Less frequently, thymine in T-T CPDs can be copied by TLS polymerases causing T→C mutations in the nTt context. While less frequent, nTt motif mutations correlate with yCn in normal human skin and in cancers (Saini et al. 2016, Saini et al. 2021).

We detected yCn MEML in dermis and epidermis, where UV-mutagenesis is expected, as well as in bladder, blood, lung, and spleen (Fig. 5A, B). Dermis samples show a high contribution of yCn MEML in CG-pair mutations (22%), second only to epidermis samples (23% in WGS, 31% in WES) (Fig 5A, Supplemental Table S7a). Sample-specific yCn MEML count range from 0 to 9,106 (median: 104, 95% confidence interval: 73-172) in dermis samples (Supplemental Table S7b). Epidermis yCn MEML range 0 to 948 in WGS samples (median: 184, 95% confidence interval: 0-499) and zero to 871 in WES samples (median: 127, 95% confidence interval: 63-252) (Fig. 5B, Supplemental Table S7b). To independently verify UV mutagenesis in tissues where yCn MEML was detected, we performed spearman correlation analyses of yCn MEML with nTt MEML (a minor motif for UV-mutagenesis) and with rCg MEML (a confounding sub-motif of nCg mutagenesis in meCpG) (Supplemental Tables S2 and S7c). There was significant positive yCn-nTt correlation in skin samples and blood, while yCn-rCg significant positive correlation was detected in lung and bladder (Fig. 5C). Removal of samples that does not have a MEML count for either motif showed a significant positive yCn-nTt correlation only in dermis (WGS) and epidermis (WES) (Fig. 5D). Based on the sub-motif analyses we concluded

that only skin samples show UV-mediated yCn mutational motif; the source of yCn MEML in other tissues can likely be attributed to spontaneous CpG deamination or other mutagenic processes. We further calculated donor-specific yCn and nTt MEML values for dermis and epidermis (Supplemental Table S7d) and assessed correlation with donor age (Supplemental Table S7e). We do not see age-dependent accumulation of UV mutational motifs in skin, except for negative correlation between nTt MEML and donor age in epidermis (Fig. 5E, Supplemental Table S7e). Since there was only small number of epidermis samples with non-zero nTt MEML, this correlation should be reconfirmed by independent studies.

Motif analyses confirm mutagenesis by APOBEC3A and APOBEC3B in bladder, breast, lung, small intestine, and liver.

APOBEC (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like) cytidine deaminases, part of the antiviral innate immunity and RNA editing systems, are also known to cause hypermutation in genomes of human cancers (Nik-Zainal et al. 2012, Roberts et al. 2013, Roberts and Gordenin 2014, Mertz et al. 2022). APOBECs cause Cytosine (C) to Uracil (U) deamination in single-stranded DNA, which can then lead to different base substitution outcomes contingent on processing by the uracil-DNA glycosylases. If no glycosylation occurs, the replicative polymerases can accurately copy the deaminated uracils as thymines and generate C→T mutations. If uracils are cleaved by glycosylation, the resulting AP-sites sites are copied by error-prone TLS polymerases which can lead to either C→T, C→G, or C→A mutations ((Saini and Gordenin 2020, Dennen et al. 2024) and references therein). Analyses of large cancer datasets have shown over representation of APOBEC-mediated C→T and C→G mutations in tCw context, with C→A mutations being barely distinguishable from random mutagenesis (Roberts et al. 2013, Mertz et al. 2022). The tCw→tTw mutational motif closely resembles COSMIC signature SBS2, while tCw→tGw is analogous to SBS13 (Alexandrov et al. 2013, Roberts et al. 2013, Alexandrov et al. 2020). APOBEC-induced tCw mutational motif enrichment, as well as mutation clusters (also

referred as kataegis) have been identified in multiple cancer types (Nik-Zainal et al. 2012, Roberts et al. 2012, Roberts et al. 2013, Sakofsky et al. 2019, Alexandrov et al. 2020). Signature analyses have also identified APOBEC mutagenesis in several healthy tissues (Moore et al. 2021, Wang et al. 2023, Yaacov et al. 2023).

We detected enrichment of tCw mutational motif across several tissues in our dataset. tCw MEML (including tCw→tTw and tCw→tGw) contributed 0% to 11% of total CG-pair mutations in the WGS samples (median: 0% , 95% confidence interval: 0%-0.3%) (Fig. 6A, Supplemental Table S8a). In WES samples, tCw contribution ranged from 0% to 7% of total CG-pair mutations (median: 0%, 95% confidence interval: 0%-0.4%) (Fig. 6A, Supplemental Table S8a). Sample-specific tCw→tTw MEML ranged from 0 to 5,589 in the WGS samples and 0-203 in WES samples (Fig. 6B, Supplemental Table S8b). tCw→tGw MEML ranged from 0 to 854 in WGS samples and 0-97 in WES samples (Fig. 6C, Supplemental Table S8b).

APOBEC-mediated mutations tend to be localized in regions of ssDNA stretches, resulting in clusters of strand-coordinated G or C mutations with inter-mutation distance of 10-10^4 bases (Nik-Zainal et al. 2012, Roberts et al. 2012, Chan and Gordenin 2015, Mertz et al. 2022). To further granulate the features of APOBEC mutagenesis, we identified the load and pattern of clustered mutations in our dataset. The percentage of samples with >0 mutation clusters ranged 12%-100% in WGS samples (median: 44%, 95% confidence interval: 33%-70%) and 4%-19% in WES samples (median: 10%, 95% confidence interval: 5%-12%) (Supplemental Table S8c). We then isolated the samples with total cluster count >0 and calculated the mean number of clusters, stratified by tissue and cluster type (Fig. 6D, Supplemental Table S8d). We see a high bias towards non-coordinated clusters, and some contribution of A- or T- coordinated clusters in all tissue type. The mean number of G- or C- coordinated clusters ranged from 0 to 8 (median range 0-1) in WGS samples; WES samples showed mean 0 to 0.7 (median range 0-1) G- or C-coordinated clusters (Fig. 6D, Supplemental Table S8d).

For the first step of identifying tissues with detectable contribution of APOBEC mutagenesis into mutation load, we defined three specific features as indicators of APOBEC mutagenesis: (i) presence of samples with statistically significant enrichment of combined tCw→tTw and tCw→tGw APOBEC motifs (Supplemental Fig. S7A), (ii) high prevalence of both tCw→tTw and tCw→tGw within G- or C-coordinated clusters in samples showing enrichment with combined APOBEC motif (Supplemental Fig. S7B), and (iii) high prevalence of G- or C-coordinated clusters (Supplemental Fig. S7C). In Supplemental Figure S7, we present additional plots for tissues that satisfied criteria for APOBEC mutagenesis. We identified multiple samples in bladder, breast, lung, liver, and small intestine tissues that showed APOBEC enrichment (Supplemental Fig. S7A). FDR-corrected statistically significant APOBEC enrichment was detected in 18% of bladder, 1% of breast, 0.3% of liver, 24% of lung, and 3% of small intestine WGS samples. In agreement with prior knowledge about APOBEC mutagenesis, there were comparable numbers of tCw→tTw and tCw→tGw mutations in G- or C- coordinated clusters of these tissues, with a small contribution of tCw→tAw mutations (Supplemental Fig. S7B). APOBEC enrichment was also detected in clusters of different sizes in these tissues (Supplemental Fig. S7C). APOBEC enrichment is expected to be less in smaller clusters and scattered mutations due to a higher chance of these clusters being formed by random mutagenesis in close vicinity. In bladder, we detect multiple clusters across all categories of G- or C- coordinated clusters with an average 3.5-fold APOBEC enrichment. Breast shows one cluster with 5 mutations and low APOBEC enrichment, while other categories of G- or C- coordinated clusters have an average 2.5-fold enrichment. We detect one cluster with 4 mutations and 9-fold APOBEC enrichment in liver samples, while enrichment in other categories of G- or C-coordinated cluster average at 3-fold. Multiple clusters across all categories of G- or C- coordinated mutations are detected in lung and small intestine samples with an average APOBEC enrichment of 3.5-fold. While some of these features were also detectable in dermis samples, APOBEC mutagenesis in G- or C-coordinated clusters was heavily skewed towards tCw→tTw MEML (Supplemental Data S1).

Since tCw motif overlaps with UV-preferred yCn motif and causes the same base substitution, we could not distinguish between the contribution of APOBEC and UV mutagenesis in skin samples. In summary, motif-centered analyses lead us to conclude that APOBEC mutagenesis is detectable in samples of normal bladder, breast, liver, lung, and small intestine tissues.

Next, we sought to identify the preference for APOBEC3A (A3A) or APOBEC3B (A3B) in tissues where APOBEC mutagenesis was identified. The A3A have been identified as greater source of cancer mutators, however contribution of A3B is also detectable (Roberts et al. 2013, Chan et al. 2015, Petljak et al. 2022). Experiments in yeast model system, and subsequent analyses of cancer datasets have revealed that hypermutation motif preference is distinguishable between A3A and A3B (Chan et al. 2015). While A3A prefers ytCa motif (APOBEC-preferred tCa motif preceded by a pyrimidine), A3B prefers rtCa motif (APOBEC-preferred tCa motif preceded by a purine). We calculated the enrichment of both ytCa (A3A-like) and rtCa (A3B-like) motifs across different mutation categories in bladder, breast, liver, lung, and small intestine tissues (Supplemental Fig. S7D, Supplemental Table S8e). Where enrichment of both ytCa and rtCa were statistically significant (Fisher's exact test), we performed Breslow-Day test of homogeneity to compare the odds ratios of the two motif enrichments (Supplemental Table S8f). Bladder, breast, lung, and small intestine showed statistically significant higher ytCa (A3A-like) enrichment than rtCa (A3B-like) across all clustered mutations (Supplemental Fig. S7D). Bladder and lung also showed preference for ytCa (A3A-like) enrichment in scattered and genome-wide mutations. Liver showed a preference for rtCa (A3B-like) enrichment; however, it was not detected to be significantly higher than ytCa (A3A-like) enrichment as both enrichment levels were relatively low compared to the other tissues. Together, our analyses revealed A3A to be the predominant APOBEC mutator in bladder, lung, breast, and small intestine healthy tissue samples.

Increased prevalence of aTn→aCn motif in various noncancer diseases

Next, we sought to compare the prevalence of the knowledge-based motifs in the healthy tissues with the diseased tissue cohorts, which included samples from 6 tissue types and 15 non-cancerous disease categories. We detected aTn motif MEML in brain, colon, liver, lung, and small intestine diseased samples (Fig. 2B, Supplemental Table S9a). Because aTn is an age-dependent motif, we categorized our cohort in three tertiles of 0-33.3, 33.4-66.7, 66.8-100 years age groups. We compared aTn MEML between healthy and diseased samples originating from the same study by Wilcoxon Rank Sum test within each age group (Fig. 7, Supplemental Table S9b). To further account for the effect of age in different disease conditions, we performed multivariable regression analyses including age and diseases as predictor variables (Supplemental Fig. S8, Supplemental Table S9c). We also performed Spearman correlation between donor age and mean donor MEML to detect clock-like accumulation of aTn MEML in diseased tissues (Supplemental Table S9d).

We did not see any statistically significant difference between healthy and diseased cohorts in the first tercile (0-33.33) of age group, regardless of the disease type. In brain, we detected statistically significant reduction of aTn MEML in samples from donors with Alzheimer's disease (Fig. 7A). Multivariable regression analyses without age categorization identified statistically significant positive effect of Alzheimer's disease on aTn MEML load, while Tourette syndrome showed significant negative effect (Supplemental Fig. S8A). Spearman correlation analyses did not reveal age dependency of aTn accumulation in any brain diseases (Supplemental Table S9d).

In colon, we identified significant positive association of aTn MEML with disease conditions within the two older age cohorts (Fig. 7B). Both Crohn's and ulcerative colitis showed a 2-fold increase in the average aTn MEML within the 33.34-66.7 age group. In the 66.7-100 age group, Crohn's samples showed 4-fold and the ulcerative colitis samples showed 3-fold increase in average aTn MEML. Regression analyses in the whole cohort of colon samples did not detect any significant association of either Crohn's or ulcerative colitis after accounting for the effect of age (Supplemental Fig. S8B). Spearman correlation analyses detected age dependent

accumulation of aTn MEML in Crohn's (Spearman's ρ= 0.5, q=0.03) and ulcerative colitis (Spearman's ρ= 0.6, q=<0.001) samples (Supplemental Table S9d).

Lung samples of donors with Chronic Obstructive Pulmonary Disorder (COPD) showed a clear disease-dependent accumulation of aTn MEML (Fig. 7C). Compared to lung samples from healthy donors, samples with COPD showed a 2.5-fold increase in the average aTn MEML within the 33.3-66.7 age group and a 4-fold increase in the 66.7-100 age group. Multivariable regression analyses identified strong positive association of COPD with aTn MEML after accounting for the effect of age (Supplemental Fig. S8C), while no clock-like activity was identified by Spearman correlation analyses (Supplemental Table S9d).

High aTn MEML was detected across all diseased liver samples compared to their healthy counterparts (Fig. 7D). Alcoholic cirrhosis samples showed 2.5-fold increase of average aTn MEML in the second tercile age group and 4-fold increase in the third tercile age group. Non-alcoholic cirrhosis samples were only available in second tercile age group and showed a 2-fold increase in average aTn MEML. Samples with Non-Alcoholic Steatohepatitis (NASH) showed 3-fold aTn MEML increase in the second tercile age group, while samples with Primary Sclerosing Cholangitis (PSC) in the same age group showed 2-fold increase (Fig. 7D). Multivariable regression analysis without age categorization showed strong positive association of all liver diseases except non-alcoholic cirrhosis, while the association of age was no longer significant (Supplemental Fig. S8D). Only donors with NASH (n=5) showed a strong age-dependent accumulation of aTn MEML (Spearman's ρ= 0.9, q=0.004) (Supplemental Table S9d).

Small intestine samples with coeliac disease showed a small but statistically significant reduction of aTn MEML in the second tercile age group, and no effect in the third tercile age group (Fig. 7E). Regression analysis identified age as the strongest predictor of aTn MEML accumulation in small intestine, with coeliac disease showing a negative but statistically

insignificant association (Supplemental Fig. S8E). No significant clock-like aTn accumulation was detected by correlation analyses in samples with coeliac disease (Supplemental Table S9d).

Increased prevalence of nCg→nTg motif in various noncancer diseases

We detected nCg MEML in all disease categories across brain, colon, epidermis, liver, lung, and small intestine tissues (Fig. 2B, Supplemental Table S10a). As previously described for healthy samples, we performed spearman correlation analyses between nCg, rCg, and yCg MEML to identify true nCg contribution (Supplemental Table S10b). We detected strong positive correlation of nCg MEML with both rCg and yCg MEML in all disease categories except psoriasis, a skin disease, which only showed correlation with yCg (Supplemental Fig. S9). We excluded psoriasis from further analyses of nCg motif.

We detected significantly high accumulation of nCg MEML in brain tissues of donors with Xeroderma Pigmentosum, while donors with Alzheimer's disease showed a reduction (Fig. 8A). In colon samples, we detected statistically significant 3-fold increase of nCg MEML in both Crohn's and ulcerative colitis diseases across all age terciles (Fig. 8B). In lung samples with COPD, we see an average 3-fold increase of nCg MEML across all age groups (Fig. 8C). In liver, a 4-fold increase of average nCg MEML was detected in donors with alcoholic cirrhosis, 4-fold increase in non-alcoholic cirrhosis, 7-fold increase in NASH, and 6-fold increase in PSC (Fig. 8D). Small intestine samples with coeliac disease showed a modest but significant increase of nCg MEML in the upper age tercile (Fig. 8E).

Multivariable regression analyses after adjusting for the effect of age showed significant positive association of increased nCg MEML with Xeroderma Pigmentosum and Alzheimer's disease in brain, ulcerative colitis in colon, COPD in lung, NASH, PSC, alcoholic and non-alcoholic cirrhosis in liver, and coeliac disease in small intestine tissues (Supplemental Fig. S8,

Supplemental Table S10d). Spearman correlation detected clock-like accumulation of nCg MEML only in donors with Tourette's Syndrome and ulcerative colitis (Supplemental Table S10e).

Discussion

In this study, we performed a meta-analysis of the somatic mutation catalogues of cancer-free normal tissues generated in dozens of independent studies to find common sources of mutations. We formed stringent statistical hypotheses to identify the enrichment of expert-curated and experimentally validated mutational motifs preferentially mutated by known mutagenic processes in the mutation catalogues of individual samples. We identified the activities of some known mutagenic mechanisms in expected tissue types, like UV-mutagenesis in skin tissues and clock-like deamination of meCpG motifs across all tissue types. We also found new, ubiquitous motif aTn→aCn associated with exposure to small epoxides and other S_N2 electrophiles across all tissues analyzed. Our previous study had identified this motif to be clock-life in multiple cancers across the PCAWG cancer catalogue (Hudson et al. 2023). Here we show that like the spontaneous deamination of methylated CpGs, mutagenesis by exposure to small epoxides accumulates throughout human lifetime and predates diseased states. We also distinguished between the preference for APOBEC3A-like and APOBEC3B-like activity in five tissue types, bladder, breast, lung, liver, and small intestine, where mutagenesis by APOBEC deaminases was detected in this current study and previously reported for both normal and cancerous tissues of these organs. These findings were enabled by the application of mutational motif-centered pipeline P-MACD accounting for prior knowledge accumulated in mechanistic research revealing the motif preference of mutagens and the molecular mechanisms of mutagenesis. The following discussion outlines the composition and limitations of the general framework of bioinformatics and statistical analyses that can be utilized for elucidating contributions of endogenous and environmental mutagenic factors into base substitution loads in human tissues.

Framework utilizing knowledge-based motif-centered analyses to dissect the sources of mutagenesis in somatic tissues of individuals

Knowledge produced by prior mechanistic research about distinct trinucleotide motif preference of a mutagenic factor to generate mutations was used for calculating statistically supported enrichment and MEML attributable to the motif-associated mutagenesis. There are only a few trinucleotide motifs so far that have been robustly associated with mutagenic factors (Table 1), but more motifs may be discovered from future efforts of the research community. Since there are only 4 bases and 96 trinucleotide motifs in DNA, a partial overlap, or even complete similarity between motifs preferred by different mutagens is inevitable. However, information derived from mechanistic research can also be used to identify non-overlapping components of motifs (i.e., sub-motifs) that are expected to be preferentially mutagenized by one out of two or more mechanisms associated with the overlapping motifs (step A, Figure 9). Motifs and sub-motifs derived from mechanistic knowledge were used to perform motif-centered analyses and generate statistically supported enrichment and MEML values from the mutation catalogs of each individual sample (step B, Figure 9). In downstream analyses (step C, Figure 9), these partially interchangeable values were used to explore correlations between different motifs and/or sub-motifs. Revealed correlations and prior mechanistic knowledge were used to resolve ambiguity stemming from motif-motif overlaps. For example, we used correlation between aTn→aCn motif and its sub-motif aTr→aCr, which does not overlap with a minor UV-motif nTt→nCt, to dissect and rule out enrichment of epoxide mutagenesis in epidermis (Figure 3C). Importantly, mutation load in epidermis was enriched for major UV-motif yCn→yTn, which in turn correlated with the minor UV-motif nTt→nCt (Figure 5C, D). In general, step of motif and sub-motif construction is aided by correlation analysis applied to motif-motif pairs chosen based on sequence overlap and prior knowledge about mutagens underlying motifs. Sample-specific values produced by motif-

centered pipeline can also be utilized in downstream statistical analyses to compare cohorts and perform correlations with the biological features of the samples and donors within cohorts.

Motif-centered analyses can identify mutagenic mechanisms operating universally or in a tissue-specific manner

The most ubiquitous motif identified in multiple samples across all tissue types was nCg→nTg associated with cytosine deamination in 5meCpG (Figure 2 and Supplemental Fig. S3). The four trinucleotides built into the nCg→nTg mutational motif (C→T mutations in aCg, cCg, gCg, and tCg) mimic the main peaks in COSMIC signature SBS1, a signature reliably extracted by different agnostic signature extraction algorithms in somatic and cancerous tissue types. Both SBS1 signature and nCg→nTg motif are age dependent, conforming to the expectation for endogenous mutational process of non-enzymatic deamination and providing additional assurance in their mechanistic assignment.

Another ubiquitous motif revealed by our analyses in healthy and diseased normal tissues was aTn→aCn associated with mutagenesis by epoxides and other S_N2 reacting electrophiles. Peaks of aTn motif closely matched the peaks of COSMIC signature SBS16 extracted from large PCAWG dataset of human cancer mutations. However, SBS16 was extracted from only two cancer types (head and neck and liver) (Alexandrov et al. 2020), while aTn→aCn motif was identified in virtually all cancer types (Hudson et al. 2023). Several studies extracted SBS16 from a limited number of somatic tissue types. SBS16 was found in neurons (Abascal et al. 2021, Ganz et al. 2024), cirrhotic liver (Brunner et al. 2019), esophagus (Moore et al. 2021), lung (Yoshida et al. 2020), and breast (Nishimura et al. 2023). Of note, SBS16 was extracted from breast tissues by only one out of three different signature extraction algorithms utilized, and was reported as a low-confidence signature by the authors (Nishimura et al. 2023). SBS16 was also not detected by MutationalPatterns package (Blokzijl et al. 2018) applied to the datasets of somatic mutations analyzed in this study (Supplemental Fig. S1 and S2). On the contrary, our motif-centered

analyses found widespread prevalence of aTn→aCn motif in all tissue types analyzed (Fig 2). In addition, motif-centered analyses utilizing sub-motif construction and motif-motif correlation separated aTn mutagenesis from confounding input of T→C mutations by UV-mutagenesis (Fig. 3C).

Framework of motif centered analysis further proved to be productive in identifying tissue-specific mutation mechanisms, as exemplified by the detection of UV and APOBEC mutagenesis. UV-associated mutational motifs yCn and nTt were used based on prior mechanistic evidence and were found only in skin (Figure 5). Sub-motif analyses allowed to rule out UV-like mutagenesis as confounded by other mechanisms in several other tissues where yCn motif was detected. Peaks corresponding to trinucleotide motifs yCn and nTt are detectable among the 96-motif spectra of UV-associated COSMIC signatures SBS7a,b and SBS7c,d, respectively, however other peaks were also present. None of SBS7 signatures were detected in somatic samples by MutationalPatterns package (Supplemental Fig. S1 and S2). Trinucleotides in APOBEC-associated motif tCw→tTw corresponds to SBS2, while tCw→tGw corresponds to SBS13 peaks. Motif-centered analyses identified APOBEC mutagenesis in bladder, liver, lung, breast, and small intestine (Figure 6, Supplemental Fig. S7), as opposed to MutationalPatterns tool which identified low contribution of SBS2 in spleen and tonsil and did not identify SBS13 in any tissue (Supplemental table S3c). Construction of tetranucleotide motifs based on prior knowledge about motif preference by APOBEC3A and APOBEC3B (Chan et al. 2015) allowed the assignment of APOBEC3A as the predominant mutator in tissues where APOBEC mutagenesis was identified (Supplemental Fig. S7), an outcome not achievable by signature analyses.

We also detected sporadic presence of some motifs in multiple tissues and diseases. Samples with enrichment of gCn→gAn motif, which is associated with acetaldehyde exposure and found primarily enriched in liver cancers (Vijayraghavan et al. 2022), were detected in multiple somatic tissue types analyzed in our study (12 out of 24; Figure 2, Supplemental table S4).

cTg→cAg motif associated with exposure to aristolochic acid (AA) was detected in healthy colon and bladder, and in both healthy and diseased samples of liver and lung (Figure 2, Supplemental table S4). SBS22a and SBS22b are associated with AA exposure and contain peaks mimicking cTg→cAg motif among other peaks. SBS22 has been extracted from cancers in upper urinary tract, liver, kidney, and from lung cancer in never-smoking East Asian individuals (Chevalier et al. 2025, Díaz-Gay et al. 2025). SBS22 was also extracted from three healthy liver tissue donors and from donors with alcohol-related liver diseases (Brunner et al. 2019, Li et al. 2021). To our knowledge, this is the first indication to AA-induced mutagenesis in normal lung tissues.

We note that the limitations of motif-centered analyses are due to the restrictive availability of knowledge about high confidence, experimentally validated motifs and the unavoidable overlap between some motifs that cannot be resolved by sub-motif approach. Signature extraction method on the other hand has unique capability to find new mutation patterns amenable for hypothesis generation. We therefore view both approaches as complementary and essential in identifying and validating sources of somatic mutations.

Motif-specific mutational burdens can serve as indicators of an individual's inherited susceptibilities and cumulative exposures over the lifespan

We explored all motifs and all datasets for potential association between somatic mutagenesis and disease, and found several diseases associated with higher prevalence of motif-specific mutation burdens caused by deamination of 5meCpG and/or exposure to small epoxides (Table 2).

Table 2. Summary of somatic mutagenesis in non-cancer diseases detected by motif-centered analysis

Tissue	Disease name	Disease specific physiology	Small epoxide associated mutagenesis (aTn→aCn)	Spontaneous deamination of meCpG (nCg→nTg)
Brain	Alzheimer's disease	Amyloid plaques and tau tangles (not fully understood)	- Decreased mutation	- Decreased mutation[1]
Brain	Xeroderma Pigmentosum[2]	DNA repair defect		- Increased mutation
Brain	Tourette's Syndrome	Neurotransmitter disruption (not fully understood)		- Clock-like[3]
Colon	Crohn's disease	Chronic inflammation; autoimmune	- Clock-like[3] - Increased mutation[4]	- Increased mutation
Colon	Ulcerative colitis	Chronic inflammation; autoimmune	- Clock-like[3] - Increased mutation[5]	- Clock-like[3] - Increased mutation
Lung	Chronic Obstructive Pulmonary Disease	Chronic inflammation	- Increased mutation	- Increased mutation
Liver	Alcoholic Cirrhosis	Chronic inflammation, fibrosis	- Increased mutation	- Increased mutation
Liver	Non-Alcoholic Cirrhosis	Chronic inflammation, fibrosis		- Increased mutation
Liver	Non-Alcoholic Steatohepatitis	Chronic inflammation, fat accumulation	- Clock-like[3] - Increased mutation	- Clock-like[3] - Increased mutation
Liver	Primary Sclerosing Cholangitis	Chronic inflammation; autoimmune	- Increased mutation	- Increased mutation
Small intestine	Coeliac	Autoimmune reaction to gluten	- Decreased mutation[6]	- Increased mutation[7]

[1] In upper age tercile of one study
[2] Non-neurological disease caused by defect in global nucleotide excision repair (NER)
[3] In the diseased cohort of the specified tissue
[4] In median age tercile
[5] In median and upper age terciles
[6] In lower age tercile
[7] In upper age tercile

Importantly, increased mutagenesis in some diseases was found for only one out of the two ubiquitous clock-like motifs which suggested an involvement of disease-specific physiological causes as opposed to a strictly age-dependent increase in mutagenesis. Several diseases associated with motif-specific mutagenesis are also associated with increased inflammation. The excess of mutations in non-neoplastic inflammatory diseases can stem from the higher rate of exposure and DNA damage compared to healthy tissues. Continued exposure to xenobiotic sources of DNA damage, like smoking in COPD patients and alcohol consumption in cirrhosis patients, increases the load of mutagenic lesions. Xenobiotic precursors can be endogenously

metabolized into reactive small epoxides by CYP2E1 and other CYP epoxygenases primarily in the liver and released back into the body (Zanger and Schwab 2013, Chen and Wang 2015). Reactive oxygen and nitrogen species (RONs) produced in inflammatory diseases can overwhelm the DNA damage repair pathways and interfere with DNA repair efficiency (Kay et al. 2019). Reactive oxygen species may further promote deamination of meCpG dinucleotides (Cao et al. 2009).

We detected a high burden of meCpG mutagenesis, but not of small epoxide motif mutagenesis, in the neurons of donors with Xeroderma Pigmentosa carrying mutations in XPA or XPD which eliminate the capacity for global nucleotide excision repair (NER) of bulky DNA lesions (Lehmann et al. 2018). However, U:G mis-pair created by meCpG deamination is a substrate for base-excision repair or for non-canonical mismatch repair, rather than for NER (Chen and Furano 2015). Future mechanistic studies may shed light on this apparent inconsistency.

The effect of age and disease on somatic mutagenesis in normal tissues can also be supplemented by the impact of individuals' genotype, which is well-established for mutagenesis in cancers (Hodgson et al. 2025). Specifically, impact of germline genotype was established on mutagenic processes in cancers that are associated with somatic mutagenesis motifs nCg (meCpG deamination) and tCw (APOBEC) revealed in this study. Germline and somatic variants in mismatch repair complex MutSα are associated with hypermutation of meCpG motifs in cancers (Flynn et al. 2024). Germline mutant alleles of meCpG binding glycosylase MBD4 also predispose to meCpG mutagenesis in cancers (Consortium 2020, Palles et al. 2022, Querido et al. 2025). Common polymorphisms in APOBEC3 gene cluster cause predisposition to cancer and exacerbate APOBEC mutation load in tumors (Nik-Zainal et al. 2014, Middlebrooks et al. 2016, Gansmo et al. 2018, Consortium 2020). As established for cancer predisposition alleles, germline polymorphisms in DNA repair genes and disease driver genes may also expediate the accumulation of somatic mutations in healthy tissues, which in turn can result in carcinogenesis

as well as non-cancer disease progression. The existence of such association between germline variants and somatic mutagenesis in normal tissues remains to be assessed.

Since motif-specific somatic mutagenesis are mechanistically linked to their sources, it allows quantification of the activities of endogenous and environmental factors that increase somatic mutation burden. Detection and monitoring of these mutagenic activities in normal tissues can prompt early diagnosis of cancer and non-cancerous diseases, which in turn may facilitate disease prevention, management, and improved treatment outcomes.

Methods

Whole genome and exome datasets

Catalogues of somatic mutations in healthy tissues from whole genome and exome sequencing mapped to GRCh38 reference genome were downloaded as VCF (Variant Call Format) files from SomaMutDB (URL: https://vijglab.einsteinmed.edu/SomaMutDB/download/) (Sun et al. 2022). A general description of VCF file format can be found in (https://docs.gdc.cancer.gov/Data/File_Formats/VCF_Format/). Mutation calls from embryonic stem cells were not included in analyses. Studies reporting mutation calls in donor groups instead of individual samples were also excluded. We also excluded one study reporting variants in endometrium, as the available VCF files only contained indel calls. Catalogues of mutations in diseased tissues from whole genome and exome sequencing mapped to GRCh38 reference genome were downloaded as VCF files from the FAQ section (question 10) of SomaMutDB (URL: https://vijglab.einsteinmed.edu/SomaMutDB/documentation/). Samples with history of carcinoma and chemotherapy were excluded. For single cell brain samples in SMSDB04 and SMSDB33, VCF files were available for the putative, high confidence somatic mutations called by linked-read analysis (LiRA) pipeline and used to estimate SNV/neuron (Lodato et al. 2017, Miller et al. 2022). Only 36 out of the 212 individual neurons with available VCF files originating from these two

studies were re-called with SCcaller to detect all mutations and were made available in the SomaMutDB database. We used the VCF files reporting the putative mutation catalogues for all 212 neurons to report mutation spectrum and perform sample-specific motif-centered analyses. Neurons from SMSDB04 that were re-analyzed in SMSDB33 were only reported once with the sample IDs from SMSDB04. Mutation calls generated in studies not included in the SomaMutDB database were collected from the supplemental materials of the cited papers (Lodato et al. 2015, Lee-Six et al. 2019, Ganz et al. 2024) and from associated dbGaP studies (Saini et al. 2016, Saini et al. 2021). Coordinates of mutation calls obtained from supplemental materials and mapped to GRCh37 reference genome were converted to GRCh38 coordinates using UCSC LiftOver tool (URL: https://genome.ucsc.edu/cgi-bin/hgLiftOver). Overlapping mutation calls were removed for clones and single cells originating from the same donor in the same tissue biospecimen. Duplicate calls were recorded once and were assigned to samples with the highest variant allele frequency where such information was available. All VCF files were converted to MAF (Mutation Annotation Format) files using the vcf2maf tool (URL: https://github.com/mskcc/vcf2maf) to perform motif-centered analysis. A general description of MAF file format can be found in (https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/).

Detection of COSMIC SBS signatures in the mutation catalogues of normal tissues

Mutation data for all samples within a tissue or disease group were pooled to generate 96-trinucleotide profiles using SigProfilerMatrixGenerator tool (Khandekar et al. 2023). Contribution of mutations in normal tissues to COSMIC SBS signatures was determined using fit_to_signatures() function of the MutationalPatterns R package (Blokzijl et al. 2018). Relative contribution of signatures was calculated and plotted using plot_contribution() function. Cosine similarity of the 96-trinucleotide profiles with known SBS signatures was identified using the cos_sim_matrix() function. Cosine matrices were generated using the pheatmap R package (Kolde 2019).

Detection of enrichment of mutations in trinucleotide motifs and calculation of minimum estimate of mutation load associated with known mutagenic processes

Trinucleotide motifs and their sub-motifs are constructed based on mechanistic knowledge and experimental validation of the mutagenic preferences of the associated mutational processes. Motif-centered analyses to find statistically significant enrichment of specific mutagenic processes were performed using the P-MACD software (URL: https://github.com/NIEHS/P-MACD) as previously described (Roberts et al. 2013, Saini et al. 2016, Saini et al. 2021, Hudson et al. 2023). Briefly, enrichment of a mutational motif was calculated by normalizing the incidence of a specific base substitution in a trinucleotide motif to the random incidence of that mutation in the background genomic context. For this calculation, context is defined as ±20 bases surrounding the mutated base to concentrate on the sequenced part of the genome capturing localized effect of lesion formation, TLS, and repair, and also to avoid epigenomic influences on mutagenesis in the larger genomic regions within each sample. Reverse complements were included in calculations and motifs were reported as pyrimidine nucleotide of a base pair. The mutated base is shown in capital, and IUPAC codes are used for ambiguous nucleotides. The formula to calculate enrichment is shown below using aTn→aCn mutational motif as an example:

$$Enrichment(aTn \rightarrow aCn) = \frac{[Mutations_{aTn \rightarrow aCn}] \times [Context_t]}{[Mutations_{T \rightarrow C}] \times [Context_{atn}]}$$

We excluded 'complex' mutations, defined as mutations ≤10bp apart, from this calculation as they may arise from the activities of the translesion polymerases and may confound the detection of mutagenic mechanisms with specific motif preferences. We performed one-sided Fisher's Exact test to statistically evaluate the overrepresentation of mutations in a motif over random mutagenesis in individual samples. Ratio of the number of mutations in motif ($Mutations_{(aTn \rightarrow aCn)}$) to mutations that are not in motif (($Mutations_{(T \rightarrow C)}$) - ($Mutations_{(aTn \rightarrow aCn)}$)) were compared to the ratio of the number of bases in the context that conformed to the motif

($Context_{(atn)}$) to bases that did not conform to the motif (($Context_{(t)}$)- ($Context_{(atn)}$)). For samples that show statistically significant enrichment after correction for multiple hypothesis using Benjamini-Hochberg method (enrichment >1, q value ≤0.05), we calculated the minimum estimate of mutation load (referred as MEML) attributable to the mutagenic process associated with the trinucleotide motif. MEML was calculated using the following formula:

$$MEML(aTn \rightarrow aCn) = \frac{[Mutations_{aTn \rightarrow aCn}] \times [Enrichment_{aTn \rightarrow aCn} - 1]}{[Enrichment_{aTn \rightarrow aCn}]}$$

MEML was recorded as zero for samples that did not show statistically significant enrichment of a mutational motif. All figures report sample-specific MEML values unless otherwise indicated.

Statistical analyses and data visualization

All statistical analyses were performed in R statistical software (version 4.4.0) via RStudio. Spearman correlation was performed for all correlation analyses using cor.test() function. Age correlations were performed with average MEML count of all samples originating from a donor. wilcox.test() function was used to perform all tests of significance, followed by correction for multiple hypothesis by Benjamini-Hochberg method using p.adjust() function. Donors in the dataset were divided in three age terciles, and tests of significance were performed for each age group within individual studies that reported both healthy and diseased cohorts to avoid batch effect and minimize the effect of aging in detecting true difference between healthy and diseased tissues. For single cell brain samples in SMSDB04 and SMSDB33, MEML counts calculated by motif-centered analyses of the putative SNVs were prorated by sample-specific multipliers to estimate genome-wide MEML load. Multipliers were calculated as

$$Multiplier = \frac{Estimated\ SNVs}{Total\ Mutations\ Identified}$$

For SMSDB04, numerator values were obtained from the '*Estimated SNVs (per autosomes)*' column and the denominator values were obtained from the '*Total Mutations Identified*' column of Supplemental_table_S5 (Lodato et al. 2017). For SMSDB33, we used values from Supplemental_table_S6 (Miller et al. 2022). For numerator of MDA-amplified available VCF files, we used values from column '*Estimated SNVs (per autosomal genome, post-filtering)*' of sheet '*Rates per Neuron (MDA)*'. For numerator of PTA-amplified available VCF files, we used values from column '*Estimated SNVs (per autosomal genome)*' of sheet '*Rates per Neuron (PTA)*'. Values from columns '*Phaseable Mutations Identified*' were used as denominators for all samples from SMSDB33.

Multivariable regression analyses were performed using lm() function using age and disease status as predictor variables and motif-specific MEML as the dependent variable. Unless otherwise indicated, all data visualizations were performed using ggplot2 R package (Wickham 2016) and Adobe Illustrator (Hoppe 2020).

Data access

All mutation calls analyzed in this study are available to download in VCF format at https://vijglab.einsteinmed.edu/SomaMutDB/download/ or in the supplementary material of the indicated papers. All data generated in this study, including numerical values underlying figures and the summary statistics are in the Supplemental Tables. P-MACD output for APOBEC cluster analyses are available in Supplemental Data S1.

Competing interest

The authors declare no competing interest.

Acknowledgements

This work was supported by the US National Institutes of Health Intramural Research Program Project Z1AES103266 to D.A.G. We are thankful to Dr. Ashley Brooks for help with vcf2maf tool, to Dr. Min Shi for consult on multivariable regression analyses, to Drs Rajula Alleva-Elango and Natalya Degtyareva for advice on manuscript, and to Lois Wyrick for help with illustrations.

Author contributions: D.A.G. – supervision; S.M.S. and D.A.G. – conceptualization, writing, review and editing; S.M.S. – data curation, data analysis, visualization; L.J.K. and Y.-C.H. – software development and testing.

Figure legends

Figure 1: Somatic mutation spectra and cosine similarity with COSMIC signatures in normal tissues. **A-B**. Median counts for each base substitution including reverse complements are shown in descending order for (**A**) healthy and (**B**) diseased tissues. Only WGS samples are shown. Median counts for WES samples are shown in Supplemental Fig. S3. **C-D**. Matrix showing cosine similarity between COSMIC signatures and 96-mutational motif profiles of healthy (**C**) and diseased (**D**) WGS normal tissues. Only the top quartile of the cosine values detected in our dataset is shown. Complete cosine values and all source data are available in Supplemental Table S3.

Figure 2: Prevalence of knowledge-based motifs in normal tissues. The percentages of samples with zero and non-zero MEML counts for the 11 known mutational motifs are shown for (**A**) healthy and (**B**) diseased tissues. Only WGS samples are shown. WES samples are shown in Supplemental Fig. S3. All source data are available in Supplemental Table S4.

Figure 3: aTn mutational motif in healthy normal cells. A. The fraction (%) of aTn→aCn MEML within total AT-pair mutations (T→A, T→C, T→G, and reverse complement) is shown on Y-axis. The tissue types are plotted on X-axis. Left panel – WGS samples; right panel – WES samples. The total aTn MEML count for a tissue is shown above the respective bars. **B**. Sample-specific

aTn→aCn MEML counts shown in jitter plots. X-axis indicates tissue types; labels above panels indicate WGS and WES groups. **C**. Correlation matrix showing aTn→aCn MEML versus aTr→aCr MEML to identify true aTn mutational motif. Spearman correlations were performed between sample-specific aTn and aTr MEML counts, including MEML=0 for either motif. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to perform correlation analyses. **D.** Correlation matrix showing donor age versus mean donor MEML of aTn and aTr motifs. Spearman correlations were performed between donor age and mean MEML count of all samples from an individual donor for the indicated tissues and mutational motifs, including MEML=0 for both motifs. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method for a given motif within each sequencing group (WGS or WES). No circles indicate not enough individual donors with mean MEML>0 for to perform correlation analyses. All source data and statistical analyses can be found in Supplemental Table S5.

Figure 4: nCg mutational motif in healthy normal cells. A. The fraction (%) of nCg→nTg MEML count within total CG-pair mutations (C→A, C→G, C→T, and reverse complement) is shown on Y-axis. The tissue types are plotted on X-axis. Left panel – WGS samples; right panel – WES samples. The total nCg MEML count for a tissue is shown above the respective bars. **B.** Sample-specific nCg→nTg MEML counts shown in jitter plots. X-axis indicates tissue types; labels above panels indicate WGS and WES groups. **C.** Correlation matrix showing nCg→nTg MEML versus rCg→rTg MEML, and nCg→nTg MEML versus yCg→yTg MEML, to identify true nCg mutational motif. Spearman correlations were performed between sample-specific nCg, rCg, and yCg MEML counts, including MEML=0 for either motif. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to

perform correlation analyses. **D.** Correlation matrix showing donor age versus mean donor MEML of nCg, rCg, and yCg motifs. Spearman correlations were performed between donor age and mean MEML count of all samples from an individual donor for the indicated tissues and mutational motifs, including MEML=0 for both motifs. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method for a given motif within each sequencing group (WGS or WES). No circles indicate not enough individual donors with mean MEML>0 for to perform correlation analyses. All source data and statistical analyses can be found in Supplemental Table S6.

Figure 5: UV mutational motifs in healthy normal cells. A. The fraction (%) of yCn→yTn MEML count in total CG-pair mutations (C→A, C→G, C→T, and reverse complement) is shown on Y-axis. The tissue types are plotted on X-axis. Left panel – WGS samples; right panel – WES samples. The total yCn MEML count for a tissue is shown above the respective bars. **B.** Sample-specific yCn→yTn MEML counts shown in jitter plots. X-axis indicates tissue types; labels above panels indicate WGS and WES groups. **C.** Spearman correlation matrix showing correlations of sample-specific values of yCn→yTn MEML versus nTt→nCt MEML, and yCn→yTn MEML versus rCg→rTg MEML to eliminate confounding effects of overlapping motifs and to support hypothesis about minor UV-motif contribution of UV-mutagenesis in tissue groups that contained samples with yCn MEML. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to perform correlation analyses. Samples with MEML=0 for either motif was included. **D**. Same as **C**, but samples with MEML=0 for either motif were excluded. **E.** Mean of nTt MEML values (including MEML=0) for samples from epidermis of each donor plotted against donor age. Only epidermis, but not dermis, showed statistically significant correlation after correction for multiple hypotheses testing (Spearman's ρ= -0.8,

q=0.03). All source data and statistical analyses for all tissue types can be found in Supplemental Table S7.

Figure 6: APOBEC tCw mutational motif in healthy normal cells. A. The fraction (%) of tCw→tTw and tCw→tGw MEML count in total CG-pair mutations (C→A, C→G, C→T, and reverse complement) is shown on Y-axis. The tissue types are plotted on X-axis. Left panel – WGS samples; right panel – WES samples. The total tCw MEML count for a tissue is shown above the respective bars. **B.** Sample-specific tCw→tTw MEML counts shown in jitter plots. X-axis indicates tissue types; labels above panels indicate WGS and WES groups. **C.** Sample-specific tCw→tGw MEML counts shown in jitter plots. X-axis indicates tissue types; labels above panels indicate WGS and WES groups. **D.** Mean number of mutation clusters for samples with total number of clusters >0 in a tissue type. All source data and statistical analyses can be found in Supplemental Table S8.

Figure 7: aTn motif in diseased tissues. aTn MEML in healthy and diseased samples are shown in a boxplot for brain **(A**), colon **(B)**, lung **(C)**, liver **(D)**, and small intestine **(E)**. X-axes indicate age groups where samples were available. Labels above panels indicate source study of the samples. Only WGS samples are shown. Asterisks in boxplots represent statistical significance (Wilcoxon Rank Sum test) compared to the healthy samples within each study and age group. *P value ≤ 0.05, **P value ≤ 0.01, ***P value ≤ 0.001. All source data and statistical analyses can be found in Supplemental Table S9.

Figure 8. nCg motif in diseased tissues. nCg MEML in healthy and diseased samples are shown in a boxplot for brain **(A)**, colon **(B)**, lung **(C)**, liver **(D)**, and small intestine **(E)**. X-axes indicate age groups where samples were available. Labels above panels indicate source study of the samples. Only WGS samples are shown. Asterisks in boxplots represent statistical significance (Wilcoxon Rank Sum test) compared to the healthy samples within each study and

age group. *P value ≤ 0.05, **P value ≤ 0.01, ***P value ≤ 0.001. All source data and statistical analyses can be found in Supplemental Table S10.

Figure 9. Framework of analyses performed in this study. A. Motifs and sub-motifs are constructed based on mechanistic knowledge about the trinucleotide preference of different mutagenic processes. **B.** Motif-centered analyses are performed on the mutation catalogue of each sample to detect enrichment and MEML attributable to the associated mutagenic process. **C.** Enrichment and MEML values are utilized in downstream analyses to find correlations with other motifs, sub-motifs, and with the features of individual samples and donors, and to perform comparisons between different cohorts.

Supplemental information

Supplemental figures

Supplemental Figure S1. Analysis of COSMIC SBS signatures in healthy tissues. Relative contributions of SBS signatures are detected in **(A)** WGS and **(B)** WES samples of healthy tissues. Source data are available in Supplemental Table S3c.

Supplemental Figure S2. Analysis of COSMIC SBS signatures in diseased tissues. Relative contributions of SBS signatures are detected in **(A)** WGS and **(B)** WES samples of different disease categories. Source data are available in Supplemental Table S3d.

Supplemental Figure S3. Somatic mutation spectra and mutational motifs in exome-sequenced samples. The median counts for each base substitution including reverse complements are shown for WES samples in healthy **(A)** and diseased **(B)** tissues. The percentages of samples with zero and non-zero MEML counts for the 11 known mutational motifs calculated for all WES samples in healthy **(C)** and diseased **(D)** tissues. Source data are available in Supplemental Tables S3 and S4.

Supplemental Figure S4. Analysis of aTn→aCn sub-motifs to detect true presence of aTn motif in healthy tissues. A. Scatter plots showing aTn motif MEML plotted against aTr motif MEML in WGS samples across different tissues. Blue line indicates line of best fit from a simple linear model. **B.** Correlation matrix showing Spearman correlation between aTn→aCn MEML and aTr→aCr MEML excluding either MEML =0 to identify true aTn mutational motif. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to perform correlation analyses. **C.** Correlation matrix showing donor age versus mean donor MEML of aTn and aTr motifs excluding either MEML=0. Spearman correlations were performed between donor age and mean MEML count of all samples from an individual donor for the indicated tissues and mutational motifs. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method for a given motif within each sequencing group (WGS or WES). No circles indicate not enough individual donors with mean MEML>0 for to perform correlation analyses. All source data and statistical analyses can be found in Supplemental Table S5.

Supplemental Figure S5. Analysis of nCg→nTg sub-motifs and their correlation with each other in healthy tissues. A. Scatter plot showing nCg→nTg MEML plotted against MEML of sub-motif rCg→rTg, the component of nCg non-overlapping to UV-motif yCn, in WGS samples across different tissues. Blue line indicates line of best fit from a simple linear model. **B.** Correlation matrix showing Spearman correlation of nCg→nTg with sub-motifs rCg→rTg (UV non-overlapping) and yCg→yTg (UV overlapping), excluding either MEML =0. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to perform correlation analyses. C. Scatter plot showing nCg→nTg MEML plotted against MEML of yCh→yTh, the component of UV-motif yCn non-overlapping to nCg, in WGS samples of dermis

and epidermis. Blue line indicates line of best fit from a simple linear model. All source data and statistical analyses can be found in Supplemental Table S6.

Supplemental Figure S6. Analysis of nCg→nTg sub-motifs and their correlations with donor age in healthy tissues. A. Scatter plot showing nCg→nTg MEML plotted against MEML of sub-motif yCg→yTg, the component of nCg overlapping to UV-motif yCn, in WGS samples across different tissues. Blue line indicates line of best fit from a simple linear model. **B.** Correlation matrix showing Spearman correlation of donor age with donor-specific mean MEML value for nCg→nTg and sub-motifs rCg→rTg (UV non-overlapping) and yCg→yTg (UV overlapping), excluding either MEML =0. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to perform correlation analyses. All source data and statistical analyses can be found in Supplemental Table S6.

Supplemental Figure S7: Tissues with multiple features indicating APOBEC mutagenesis. A. Fold-enrichment of APOBEC mutagenesis motif over the expected occurrence of random mutagenesis in the indicated tissues. Samples are categorized into color coded bins based on their FDR corrected q-values (Benjamini-Hochberg) and fold enrichment values. Fold enrichment bin sizes are increments of 1 unit. All samples displaying a q-value >0.05 are placed in one bin (black) regardless of the fold APOBEC enrichment in the sample. The maximum fold enrichment for each bin is indicated in the figure legend, with the number of samples in its category shown in parentheses. **B.** The number of mutations categorized into three possible types of substitutions at the tCw motif (complementary mutations included) within C- or G-coordinated clusters in the indicated tissues. **C.** Fold enrichment of the APOBEC motif mutations in clusters of different sizes of two categories – C- or G-coordinated, non-coordinated, as well as in non-clustered (scattered) mutations in the indicated tissues. Values above the bars display the total number of mutation

clusters in each class and in parentheses the total numbers of APOBEC motif mutations in a category. **D.** Enrichment of ytCa (APOBEC3A-like) motif and rtCa (APOBEC3B-like) motif among genome-wide, scattered, clustered, and G- or C- coordinated clustered mutations in the indicated tissues. The blue dashed vertical line indicates level of no enrichment. Where enrichment of both ytCa and rtCa motifs were statistically significant (Fisher's Exact P-value ≤0.05), Breslow-Day test of homogeneity was performed to compare the odds ratios of the two motif enrichments. Asterisks indicate significant difference at P-value ≤0.05. Source data for plots **A**, **B**, and **C** are available in Supplemental Data S1. Source data and statistical analyses for ytCa and rtCa enrichment **(D)** can be found in Supplemental Table S8e and S8f.

Supplemental Figure S8. Multivariable regression analyses of aTn and nCg motifs in non-cancer diseases. *β*-coefficients from multivariable regression analyses using donor age and disease status as predictor variables are plotted for brain **(A)**, liver **(B)**, colon **(C)**, lung **(D)**, and small intestine **(E)**. *P value ≤ 0.05, **P value ≤ 0.01, ***P value ≤ 0.001. Source data are available in Supplemental Tables S9c and S10d.

Supplemental Figure S9. Analyses of sub-motifs to detect true spontaneous meCpG cytosine deamination-mediated nCg→nTg in diseased tissues. Correlation matrix showing Spearman correlation of nCg→nTg with sub-motifs rCg→rTg (UV non-overlapping) and yCg→yTg (UV overlapping), including MEML = 0 **(A)** and excluding pairs with either member's MEML =0 **(B)** in different disease categories. Grey circles indicate P value >0.05 after correction for multiple hypothesis using Benjamini-Hochberg method within each sequencing group (WGS or WES). No circles indicate not enough samples with MEML>0 for a tissue type to perform correlation analyses. Source data are available in Supplemental Table S10b.

Supplemental Tables

Supplemental Table S1: Origins of mutation calls for samples analyzed in this study.

Supplemental Table S2: Sub-motifs of knowledge-based motifs with overlapping and non-overlapping components.

Supplemental Table S3: Somatic mutation spectra and mutational motifs in normal cells.

Supplemental Table S4: Prevalence of knowledge-based motifs in normal tissues

Supplemental Table S5: aTn mutational motif in normal healthy cells.

Supplemental Table S6: nCg mutational motif in normal healthy cells.

Supplemental Table S7: UV mutational motif in normal cells.

Supplemental Table S8: APOBEC mutational motif in healthy normal tissues.

Supplemental Table S9: Prevalence of aTn motif in diseased tissues.

Supplemental Table S10: Prevalence of nCg mutational motif in diseased samples.

Supplemental Data

Supplemental Data S1: Raw output of P-MACD motif and cluster analyses to detect APOBEC mutagenesis.

References

Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, Russell AJC, Alcantara RE, Baez-Ortega A, Wang Y, et al. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature* **593**: 405-410. doi: 10.1038/s41586-021-03477-4.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* **578**: 94-101. doi: 10.1038/s41586-020-1943-3.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. (2013). Signatures of mutational processes in human cancer. *Nature* **500**: 415-421. doi: 10.1038/nature12477.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ and Stratton MR (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**: 246-259. doi: 10.1016/j.celrep.2012.12.008.

Ames BN, Durston WE, Yamasaki E and Lee FD (1973). Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. *Proc Natl Acad Sci U S A* **70**: 2281-2285. doi: 10.1073/pnas.70.8.2281.

Blokzijl F, Janssen R, van Boxtel R and Cuppen E (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Medicine* **10**: 33. doi: 10.1186/s13073-018-0539-0.

Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, Sanders MA, Ellis P, Alder C, Hooks Y, et al. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**: 538-542. doi: 10.1038/s41586-019-1670-9.

Cao H, Jiang Y and Wang Y (2009). Kinetics of deamination and Cu(II)/H 2 O 2 /Ascorbate-induced formation of 5-methylcytosine glycol at CpG sites in duplex DNA. *Nucleic Acids Res* **37**: 6635-6643. doi: 10.1093/nar/gkp615.

Chan K and Gordenin DA (2015). Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu Rev Genet* **49**: 243-267. doi: 10.1146/annurev-genet-112414-054714.

Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, Kim J, Kwiatkowski DJ, Fargo DC, Mieczkowski PA, et al. (2015). An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* **47**: 1067-1072. doi: 10.1038/ng.3378.

Chawanthayatham S, Valentine CC, Fedeles BI, Fox EJ, Loeb LA, Levine SS, Slocum SL, Wogan GN, Croy RG and Essigmann JM (2017). Mutational spectra of aflatoxin B₁

in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci U S A* **114**: E3101-E3109. doi: doi:10.1073/pnas.1700759114.

Chen C and Wang DW (2015). Chapter Seven - Cytochrome P450-CYP2 Family-Epoxygenase Role in Inflammation and Cancer. Advances in Pharmacology. J. P. Hardwick, Academic Press. **74:** 193-221.

Chen J and Furano AV (2015). Breaking bad: The mutagenic effect of DNA repair. *DNA Repair* **32**: 43-51. doi: 10.1016/j.dnarep.2015.04.012.

Chevalier A, Guo T, Gurevich NQ, Xu J, Yajima M and Campbell JD (2025). Characterization of Mutational Signatures in Tumors from a Large Chinese Population. *Cancer Research Communications* **5**: 1466-1476. doi: 10.1158/2767-9764.CRC-24-0496.

Consortium ITP-CAoWG (2020). Pan-cancer analysis of whole genomes. *Nature* **578**: 82-93. doi: 10.1038/s41586-020-1969-6.

Degtyareva NP, Placentra VC, Gabel SA, Klimczak LJ, Gordenin DA, Wagner BA, Buettner GR, Mueller GA, Smirnova TI and Doetsch PW (2023). Changes in metabolic landscapes shape divergent but distinct mutational signatures and cytotoxic consequences of redox stress. *Nucleic Acids Res* **51**: 5056-5072. doi: 10.1093/nar/gkad305.

Dennen MS, Kockler ZW, Roberts SA, Burkholder AB, Klimczak LJ and Gordenin DA (2024). Hypomorphic mutation in the large subunit of replication protein A affects mutagenesis by human APOBEC cytidine deaminases in yeast. *G3 Genes|Genomes|Genetics* **14**: jkae196. doi: 10.1093/g3journal/jkae196.

Díaz-Gay M, Zhang T, Hoang PH, Leduc C, Baine MK, Travis WD, Sholl LM, Joubert P, Khandekar A, Zhao W, et al. (2025). The mutagenic forces shaping the genomes of lung cancer in never smokers. *Nature* **644**: 133-144. doi: 10.1038/s41586-025-09219-0.

Erickson RP (2003). Somatic gene mutation and human disease other than cancer. *Mutat Res* **543**: 125-136. doi: 10.1016/s1383-5742(03)00010-3.

Erickson RP (2014). Recent advances in the study of somatic mosaicism and diseases other than cancer. *Curr Opin Genet Dev* **26**: 73-78. doi: 10.1016/j.gde.2014.06.001.

Erickson RP (2016). The importance of de novo mutations for pediatric neurological disease--It is not all in utero or birth trauma. *Mutat Res Rev Mutat Res* **767**: 42-58. doi: 10.1016/j.mrrev.2015.12.002.

Evrony GD, Hinch AG and Luo C (2021). Applications of Single-Cell DNA Sequencing. *Annu Rev Genomics Hum Genet* **22**: 171-197. doi: 10.1146/annurev-genom-111320-090436.

Flynn A, Waszak SM and Weischenfeldt J (2024). Somatic CpG hypermutation is associated with mismatch repair deficiency in cancer. *Mol Syst Biol* **20**: 1006-1024-1024. doi: https://doi.org/10.1038/s44320-024-00054-5.

Gansmo LB, Romundstad P, Hveem K, Vatten L, Nik-Zainal S, Lønning PE and Knappskog S (2018). APOBEC3A/B deletion polymorphism and cancer risk. *Carcinogenesis* **39**: 118-124. doi: 10.1093/carcin/bgx131.

Ganz J, Luquette LJ, Bizzotto S, Miller MB, Zhou Z, Bohrson CL, Jin H, Tran AV, Viswanadham VV, McDonough G, et al. (2024). Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. *Cell* **187**: 1955-1970.e1923. doi: https://doi.org/10.1016/j.cell.2024.02.025.

Hodgson SV, Foulkes WD, Maher ER and Turnbull C (2025). Inherited Susceptibility to Cancer: Past, Present and Future. *Annals of Human Genetics* **89**: 354-365. doi: https://doi.org/10.1111/ahg.70013.

Hoppe J (2020). *Adobe Illustrator CC : a complete course and compendium of features*. San Rafael, CA, Rocky Nook, Inc.

Hudson KM, Klimczak LJ, Sterling JF, Burkholder AB, Kazanov MD, Saini N, Mieczkowski PA and Gordenin DA (2023). Glycidamide-induced hypermutation in yeast single-stranded DNA reveals a ubiquitous clock-like mutational motif in humans. *Nucleic Acids Res* **51**: 9075-9100. doi: 10.1093/nar/gkad611.

Inman GJ, Wang J, Nagano A, Alexandrov LB, Purdie KJ, Taylor RG, Sherwood V, Thomson J, Hogan S, Spender LC, et al. (2018). The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. *Nature Communications* **9**: 3667. doi: 10.1038/s41467-018-06027-1.

Kay J, Thadhani E, Samson L and Engelward B (2019). Inflammation-induced DNA damage, mutations and cancer. *DNA Repair* **83**: 102673. doi: https://doi.org/10.1016/j.dnarep.2019.102673.

Khandekar A, Vangara R, Barnes M, Díaz-Gay M, Abbasi A, Bergstrom EN, Steele CD, Pillay N and Alexandrov LB (2023). Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. *BMC Genomics* **24**: 469. doi: 10.1186/s12864-023-09584-y.

Kolde R (2019). Pheatmap: pretty heatmaps. *R package version* **1**: 726.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214-218. doi: 10.1038/nature12213.

Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, Georgakopoulos N, Torrente F, Noorani A, Goddard M, et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**: 532-537. doi: 10.1038/s41586-019-1672-7.

Lehmann J, Seebode C, Martens MC and Emmert S (2018). Xeroderma Pigmentosum - Facts and Perspectives. *Anticancer Res* **38**: 1159-1164. doi: 10.21873/anticanres.12335.

Li R, Di L, Li J, Fan W, Liu Y, Guo W, Liu W, Liu L, Li Q, Chen L, et al. (2021). A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**: 398-403. doi: 10.1038/s41586-021-03836-1.

Lodato MA, Rodin RE, Bohrson CL, Coulter ME, Barton AR, Kwon M, Sherman MA, Vitzthum CM, Luquette LJ, Yandava C, et al. (2017). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* doi: 10.1126/science.aao4426.

Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, D'Gama AM, Cai X, et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**: 94-98. doi: 10.1126/science.aab1785.

Loeb LA, Springgate CF and Battula N (1974). Errors in DNA replication as a basis of malignant changes. *Cancer Res* **34**: 2311-2321.

Martincorena I and Campbell PJ (2015). Somatic mutation in cancer and normal cells. *Science* **349**: 1483-1489. doi: 10.1126/science.aab4082.

Maslov AY and Vijg J (2023). Somatic mutation burden in relation to aging and functional life span: implications for cellular reprogramming and rejuvenation. *Curr Opin Genet Dev* **83**: 102132. doi: 10.1016/j.gde.2023.102132.

Mertz TM, Collins CD, Dennis M, Coxon M and Roberts SA (2022). APOBEC-Induced Mutagenesis in Cancer. *Annu Rev Genet* doi: 10.1146/annurev-genet-072920-035840.

Middlebrooks CD, Banday AR, Matsuda K, Udquim KI, Onabajo OO, Paquin A, Figueroa JD, Zhu B, Koutros S, Kubo M, et al. (2016). Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat Genet* **48**: 1330-1338. doi: 10.1038/ng.3670.

Miller MB, Huang AY, Kim J, Zhou Z, Kirkham SL, Maury EA, Ziegenfuss JS, Reed HC, Neil JE, Rento L, et al. (2022). Somatic genomic changes in single Alzheimer's disease neurons. *Nature* doi: 10.1038/s41586-022-04640-1.

Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, Oliver TRW, Leongamornlert D, Ellis P, Noorani A, et al. (2021). The mutational landscape of human somatic and germline cells. *Nature* **597**: 381-386. doi: 10.1038/s41586-021-03822-7.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**: 979-993. doi: 10.1016/j.cell.2012.04.024.

Nik-Zainal S, Wedge DC, Alexandrov LB, Petljak M, Butler AP, Bolli N, Davies HR, Knappskog S, Martin S, Papaemmanuil E, et al. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet* **46**: 487-491. doi: 10.1038/ng.2955.

Nishimura T, Kakiuchi N, Yoshida K, Sakurai T, Kataoka TR, Kondoh E, Chigusa Y, Kawai M, Sawada M, Inoue T, et al. (2023). Evolutionary histories of breast cancer and related clones. *Nature* **620**: 607-614. doi: 10.1038/s41586-023-06333-9.

Palles C, West HD, Chew E, Galavotti S, Flensburg C, Grolleman JE, Jansen EAM, Curley H, Chegwidden L, Arbe-Barnes EH, et al. (2022). Germline MBD4 deficiency causes a multi-tumor predisposition syndrome. *Am J Hum Genet* **109**: 953-960. doi: 10.1016/j.ajhg.2022.03.018.

Peng W and Shaw BR (1996). Accelerated deamination of cytosine residues in UV-induced cyclobutane pyrimidine dimers leads to CC-->TT transitions. *Biochemistry* **35**: 10172-10181. doi: 10.1021/bi960001x.

Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, Dawson KJ, Ju YS, Iorio F, Tubio JMC, et al. (2019). Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**: 1282-1294 e1220. doi: 10.1016/j.cell.2019.02.012.

Petljak M, Dananberg A, Chu K, Bergstrom EN, Striepen J, von Morgen P, Chen Y, Shah H, Sale JE, Alexandrov LB, et al. (2022). Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* **607**: 799-807. doi: 10.1038/s41586-022-04972-y.

Pfeifer GP (2006). Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol* **301**: 259-281.

Querido I, Pinto C, Arinto P, Brandão A, Santos C, Pinheiro M, Guerra J, Silva J, Peixoto A and Teixeira MR (2025). A Novel Deleterious Variant and a Founder Effect in Four New Families of MBD4-Associated Neoplasia Syndrome Recruited Over a Period of 20 Years. *Clin Genet* doi: 10.1111/cge.70014.

Roberts SA and Gordenin DA (2014). Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer* **14**: 786-800. doi: 10.1038/nrc3816.

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**: 970-976. doi: 10.1038/ng.2702.

Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA, et al. (2012). Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46**: 424-435. doi: 10.1016/j.molcel.2012.03.030.

Rosenquist TA and Grollman AP (2016). Mutational signature of aristolochic acid: Clue to the recognition of a global disease. *DNA Repair (Amst)* **44**: 205-211. doi: 10.1016/j.dnarep.2016.05.027.

Saini N, Giacobone CK, Klimczak LJ, Papas BN, Burkholder AB, Li JL, Fargo DC, Bai R, Gerrish K, Innes CL, et al. (2021). UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin. *PLoS Genet* **17**: e1009302. doi: 10.1371/journal.pgen.1009302.

Saini N and Gordenin DA (2018). Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. *Environ Mol Mutagen* **59**: 672-686. doi: 10.1002/em.22215.

Saini N and Gordenin DA (2020). Hypermutation in single-stranded DNA. *DNA Repair (Amst)* **91-92**: 102868. doi: 10.1016/j.dnarep.2020.102868.

Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, Fargo DC, Boyer JC, Kaufmann WK, Taylor JA, et al. (2016). The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. *PLoS Genet* **12**: e1006385. doi: 10.1371/journal.pgen.1006385.

Saini N, Sterling JF, Sakofsky CJ, Giacobone CK, Klimczak LJ, Burkholder AB, Malc EP, Mieczkowski PA and Gordenin DA (2020). Mutation signatures specific to DNA alkylating agents in yeast and cancers. *Nucleic Acids Res* **48**: 3692-3707. doi: 10.1093/nar/gkaa150.

Sakofsky CJ, Saini N, Klimczak LJ, Chan K, Malc EP, Mieczkowski PA, Burkholder AB, Fargo D and Gordenin DA (2019). Repair of multiple simultaneous double-strand breaks causes bursts of genome-wide clustered hypermutation. *PLoS Biol* **17**: e3000464. doi: 10.1371/journal.pbio.3000464.

Sun S, Wang Y, Maslov AY, Dong X and Vijg J (2022). SomaMutDB: a database of somatic mutations in normal human tissues. *Nucleic Acids Res* **50**: D1100-d1108. doi: 10.1093/nar/gkab914.

Vijayraghavan S, Porcher L, Mieczkowski PA and Saini N (2022). Acetaldehyde makes a distinct mutation signature in single-stranded DNA. *Nucleic Acids Res* doi: 10.1093/nar/gkac570.

Wang Y, Robinson PS, Coorens THH, Moore L, Lee-Six H, Noorani A, Sanders MA, Jung H, Katainen R, Heuschkel R, et al. (2023). APOBEC mutagenesis is a common process in normal human small intestine. *Nat Genet* doi: 10.1038/s41588-022-01296-5.

Wickham H (2016). Data analysis. *ggplot2: elegant graphics for data analysis*, Springer**:** 189-201.

Yaacov A, Rosenberg S and Simon I (2023). Mutational signatures association with replication timing in normal cells reveals similarities and differences with matched cancer tissues. *Sci Rep* **13**: 7833. doi: 10.1038/s41598-023-34631-9.

Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, Beal K, Menzies A, Millar FR, Anderson E, et al. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* doi: 10.1038/s41586-020-1961-1.

You Y-H, Lee D-H, Yoon J-H, Nakajima S, Yasui A and Pfeifer GP (2001). Cyclobutane Pyrimidine Dimers Are Responsible for the Vast Majority of Mutations Induced by UVB Irradiation in Mammalian Cells*. *J Biol Chem* **276**: 44688-44694. doi: https://doi.org/10.1074/jbc.M107696200.

Zanger UM and Schwab M (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther* **138**: 103-141. doi: 10.1016/j.pharmthera.2012.12.007.

Zhang L, Lee M, Maslov AY, Montagna C, Vijg J and Dong X (2024). Analyzing somatic mutations by single-cell whole-genome sequencing. *Nat Protoc* **19**: 487-516. doi: 10.1038/s41596-023-00914-8.

Zhang T, Sang J, Hoang PH, Zhao W, Rosenbaum J, Johnson KE, Klimczak LJ, McElderry J, Klein A, Wirth C, et al. (2025). APOBEC affects tumor evolution and age at onset of lung cancer in smokers. *Nat Commun* **16**: 4711. doi: 10.1038/s41467-025-59923-8.

Figure 1

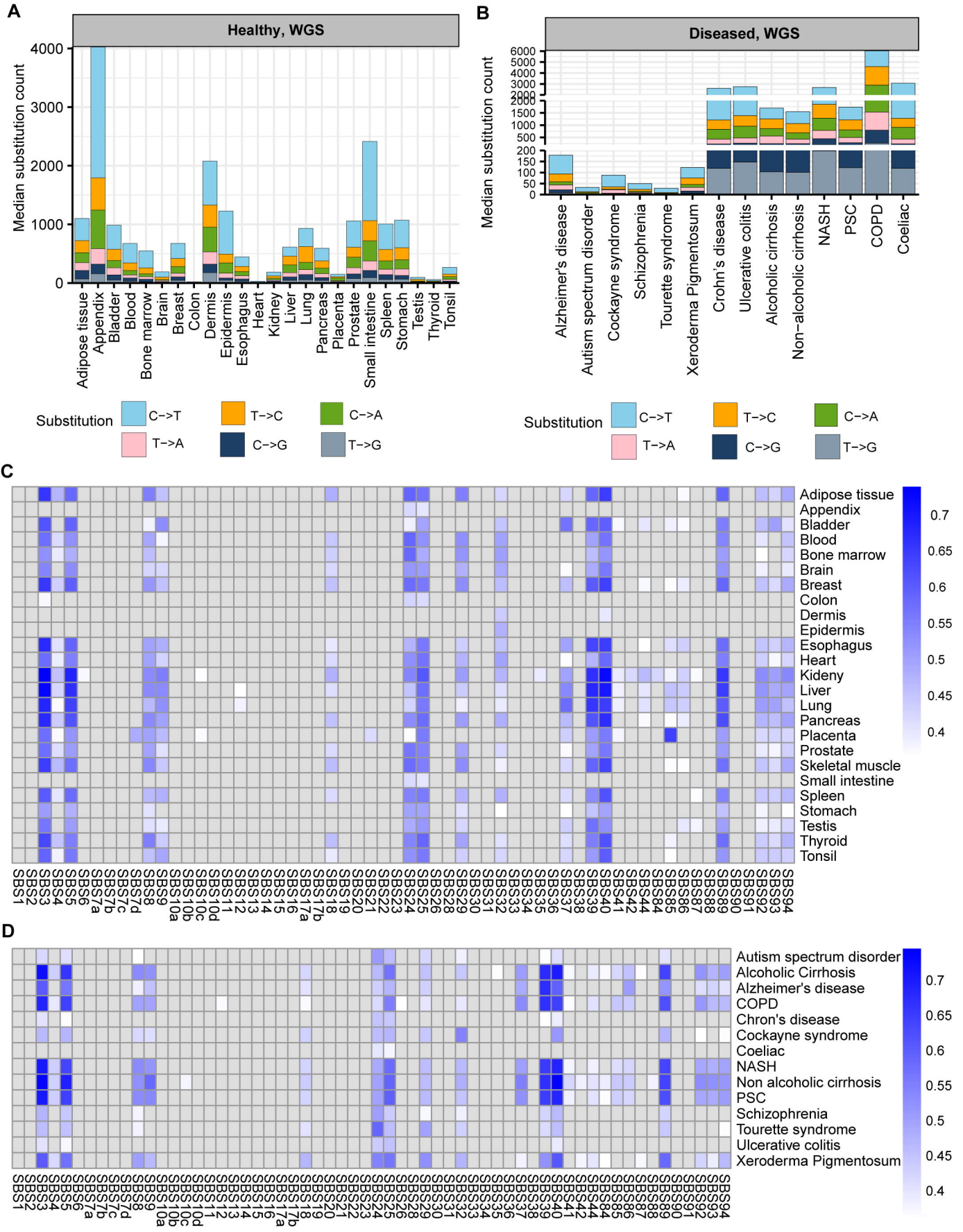

A
Healthy, WGS
Median substitution count
4000
3000
2000
1000
0
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kidney
Liver
Lung
Pancreas
Placenta
Prostate
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
Substitution
C->T
T->C
C->A
T->A
C->G
T->G
B
Diseased, WGS
Median substitution count
6000
5000
4000
3000
2000
1500
1000
500
200
150
100
50
0
Alzheimer's disease
Autism spectrum disorder
Cockayne syndrome
Schizophrenia
Tourette syndrome
Xeroderma Pigmentosum
Crohn's disease
Ulcerative colitis
Alcoholic cirrhosis
Non-alcoholic cirrhosis
NASH
PSC
COPD
Coeliac
Substitution
C->T
T->C
C->A
T->A
C->G
T->G
C
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kideny
Liver
Lung
Pancreas
Placenta
Prostate
Skeletal muscle
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
0.7
0.65
0.6
0.55
0.5
0.45
0.4
SBS1 SBS2 SBS3 SBS4 SBS5 SBS6 SBS7a SBS7b SBS7c SBS7d SBS8 SBS9 SBS10a SBS10b SBS10c SBS10d SBS11 SBS12 SBS13 SBS14 SBS15 SBS16 SBS17a SBS17b SBS18 SBS19 SBS20 SBS21 SBS22 SBS23 SBS24 SBS25 SBS26 SBS28 SBS29 SBS30 SBS31 SBS32 SBS33 SBS34 SBS35 SBS36 SBS37 SBS38 SBS39 SBS40 SBS41 SBS42 SBS44 SBS45 SBS46 SBS47 SBS84 SBS85 SBS86 SBS87 SBS88 SBS89 SBS90 SBS91 SBS92 SBS93 SBS94
D
Autism spectrum disorder
Alcoholic Cirrhosis
Alzheimer's disease
COPD
Chron's disease
Cockayne syndrome
Coeliac
NASH
Non alcoholic cirrhosis
PSC
Schizophrenia
Tourette syndrome
Ulcerative colitis
Xeroderma Pigmentosum
0.7
0.65
0.6
0.55
0.5
0.45
0.4
SBS1 SBS2 SBS3 SBS4 SBS5 SBS6 SBS7a SBS7b SBS7c SBS7d SBS8 SBS9 SBS10a SBS10b SBS10c SBS10d SBS11 SBS12 SBS13 SBS14 SBS15 SBS16 SBS17a SBS17b SBS18 SBS19 SBS20 SBS21 SBS22 SBS23 SBS24 SBS25 SBS26 SBS28 SBS29 SBS30 SBS31 SBS32 SBS33 SBS34 SBS35 SBS36 SBS37 SBS38 SBS39 SBS40 SBS41 SBS42 SBS44 SBS45 SBS46 SBS47 SBS84 SBS85 SBS86 SBS87 SBS88 SBS89 SBS90 SBS91 SBS92 SBS93 SBS94

Figure 2

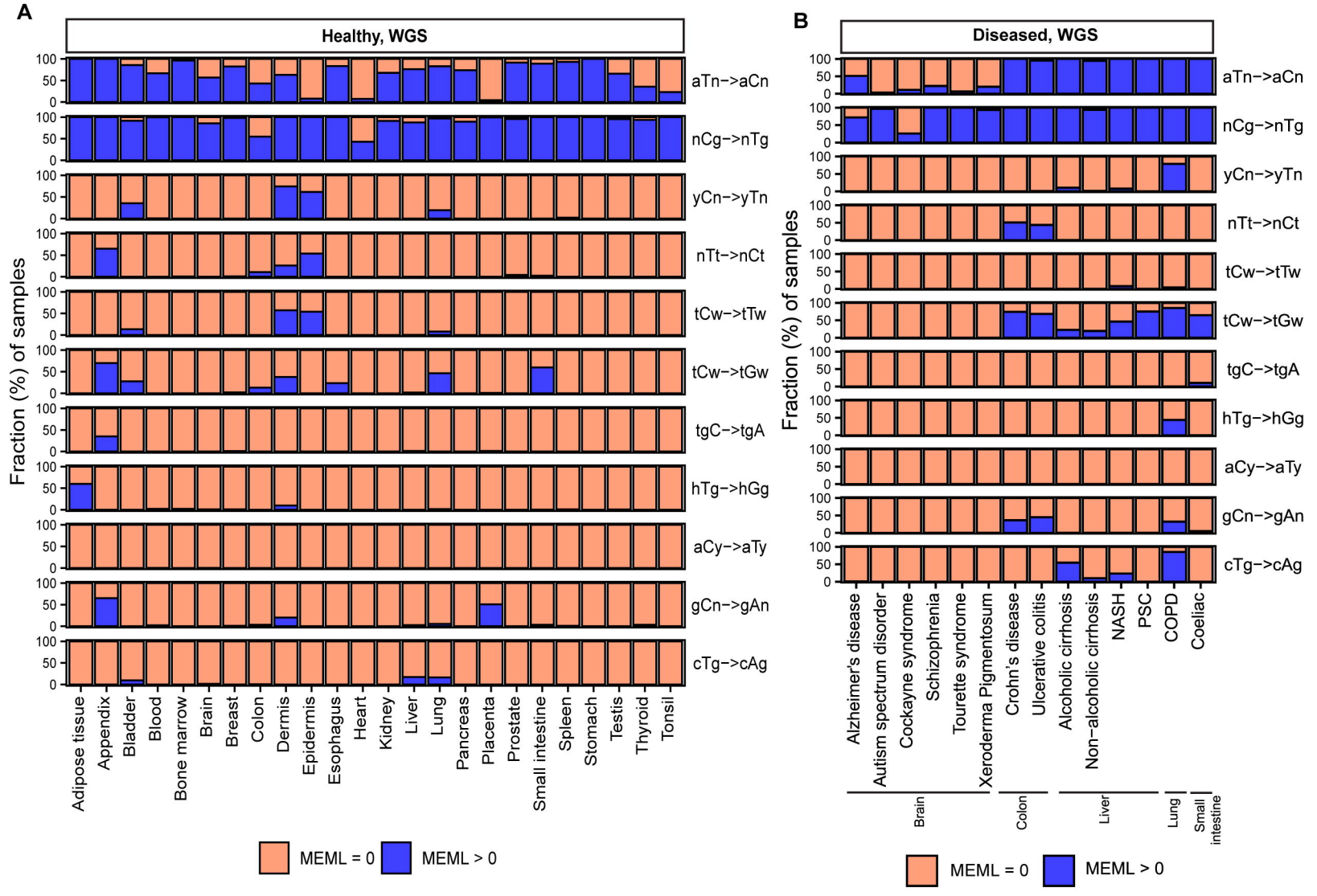

A
Healthy, WGS
Fraction (%) of samples
100
50
0
aTn->aCn
nCg->nTg
yCn->yTn
nTt->nCt
tCw->tTw
tCw->tGw
tgC->tgA
hTg->hGg
aCy->aTy
gCn->gAn
cTg->cAg
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kidney
Liver
Lung
Pancreas
Placenta
Prostate
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
MEML = 0
MEML > 0
B
Diseased, WGS
Fraction (%) of samples
Alzheimer's disease
Autism spectrum disorder
Cockayne syndrome
Schizophrenia
Tourette syndrome
Xeroderma Pigmentosum
Crohn's disease
Ulcerative colitis
Alcoholic cirrhosis
Non−alcoholic cirrhosis
NASH
PSC
COPD
Coeliac
Brain
Colon
Liver
Lung
Small intestine
MEML = 0
MEML > 0

Figure 3

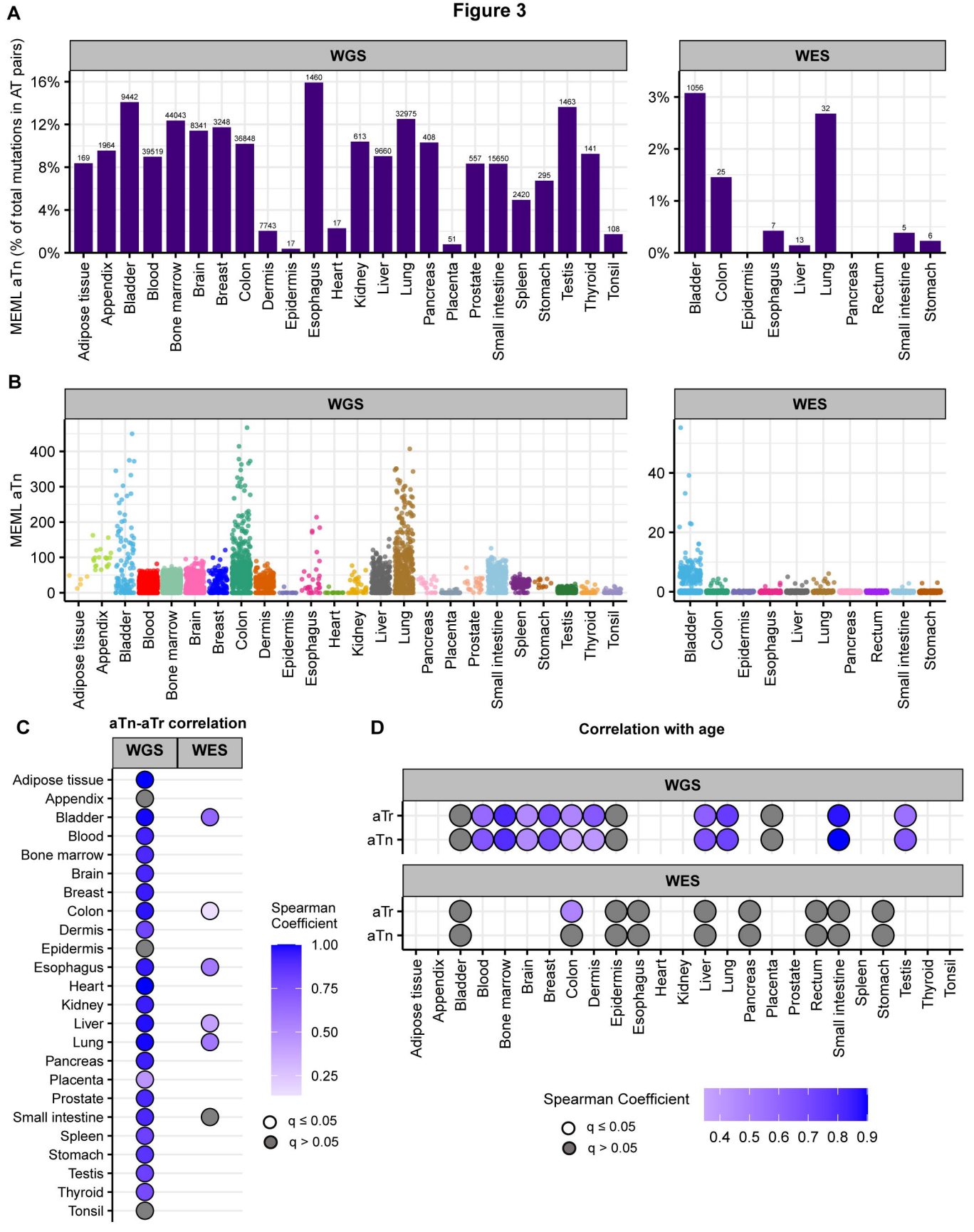

A
WGS
WES
MEML aTn (% of total mutations in AT pairs)
16%
12%
8%
4%
0%
3%
2%
1%
169
1964
9442
39519
44043
8341
3248
36848
7743
17
1460
17
613
9660
32975
408
51
557
15650
2420
295
1463
141
108
1056
25
7
13
32
5
6
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kidney
Liver
Lung
Pancreas
Placenta
Prostate
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
Rectum
B
MEML aTn
400
300
200
100
0
40
20
C
aTn-aTr correlation
Spearman Coefficient
1.00
0.75
0.50
0.25
q ≤ 0.05
q > 0.05
D
Correlation with age
aTr
aTn
Spearman Coefficient
0.4
0.5
0.6
0.7
0.8
0.9

Figure 4

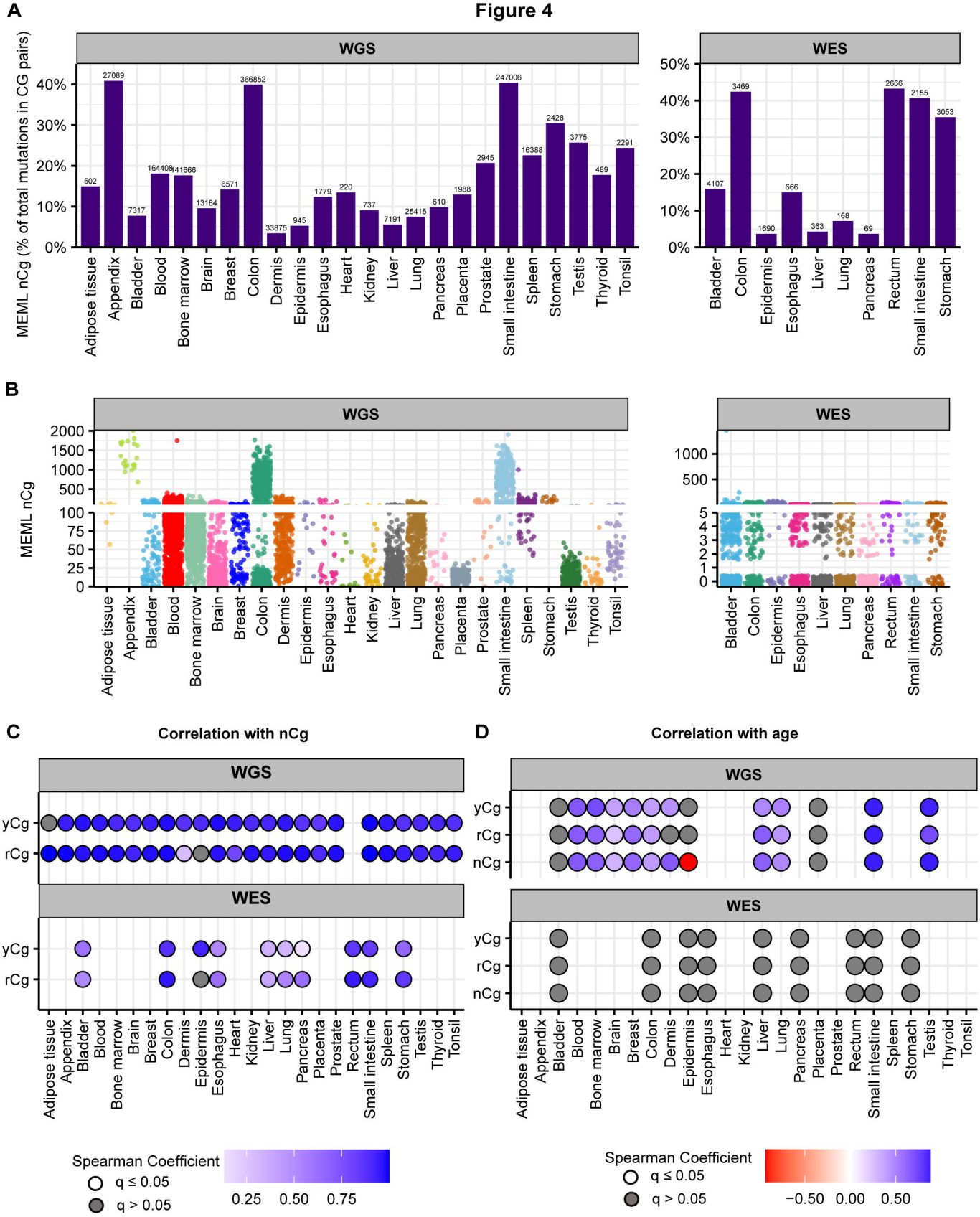

A
WGS
WES
MEML nCg (% of total mutations in CG pairs)
0%
10%
20%
30%
40%
50%
502
27089
7317
164408
141666
13184
6571
366852
33875
945
1779
220
737
7191
25415
610
1988
2945
247006
16388
2428
3775
489
2291
4107
3469
1690
666
363
168
69
2666
2155
3053
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kidney
Liver
Lung
Pancreas
Placenta
Prostate
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
Rectum
B
MEML nCg
C
Correlation with nCg
yCg
rCg
D
Correlation with age
nCg
Spearman Coefficient
q ≤ 0.05
q > 0.05
0.25
0.50
0.75
−0.50
0.00
0.50

Figure 5

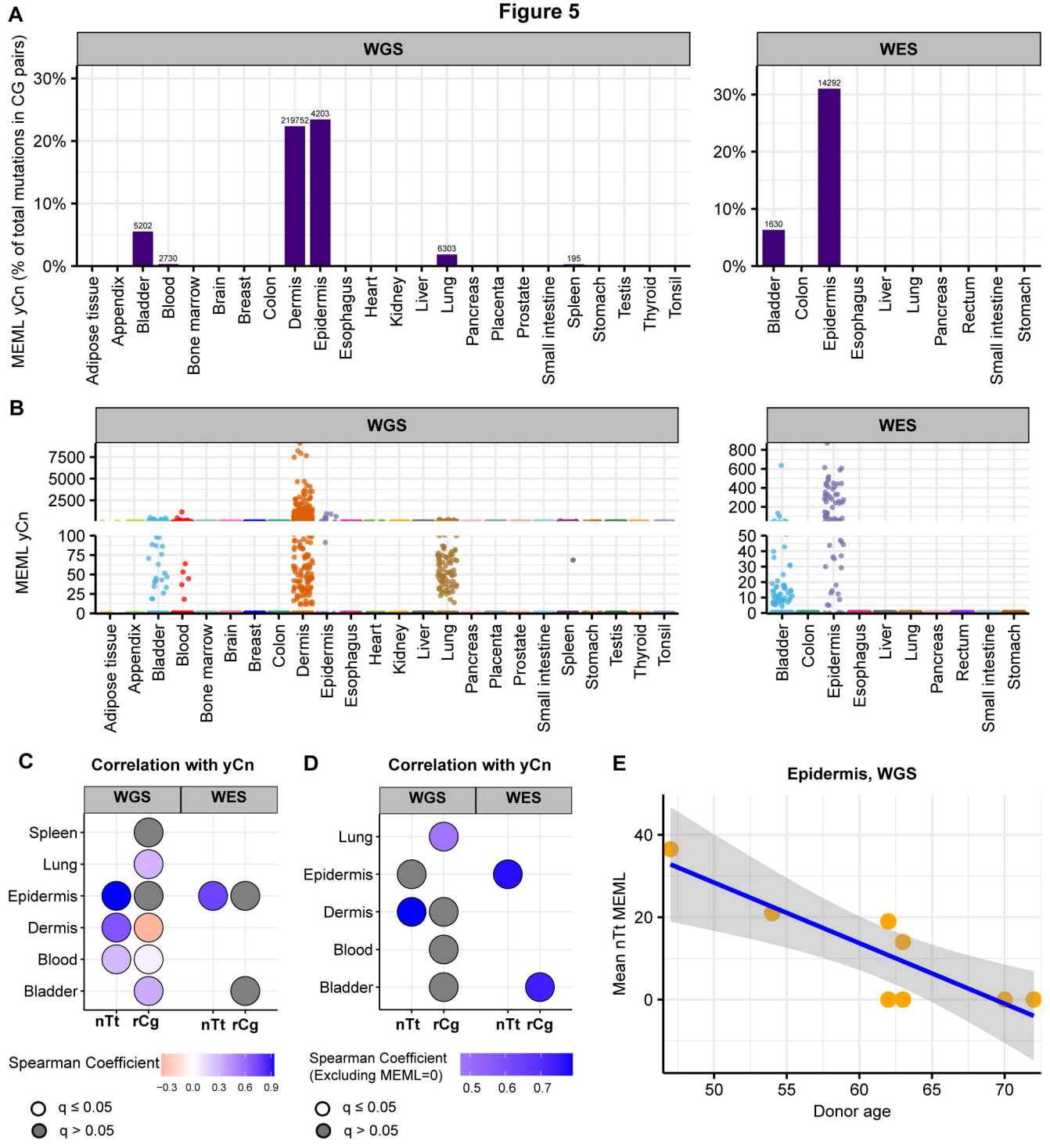

A
WGS
WES
MEML yCn (% of total mutations in CG pairs)
30%
20%
10%
0%
5202
2730
219752
4203
6303
195
1630
14292
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kidney
Liver
Lung
Pancreas
Placenta
Prostate
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
Rectum
B
MEML yCn
7500
5000
2500
100
75
50
25
0
800
600
400
200
50
40
30
20
10
C
Correlation with yCn
Spleen
Lung
Epidermis
Dermis
Blood
Bladder
nTt
rCg
Spearman Coefficient
−0.3 0.0 0.3 0.6 0.9
q ≤ 0.05
q > 0.05
D
Correlation with yCn
Spearman Coefficient
(Excluding MEML=0)
0.5 0.6 0.7
E
Epidermis, WGS
Mean nTt MEML
Donor age
50
55
60
65
70

Figure 6

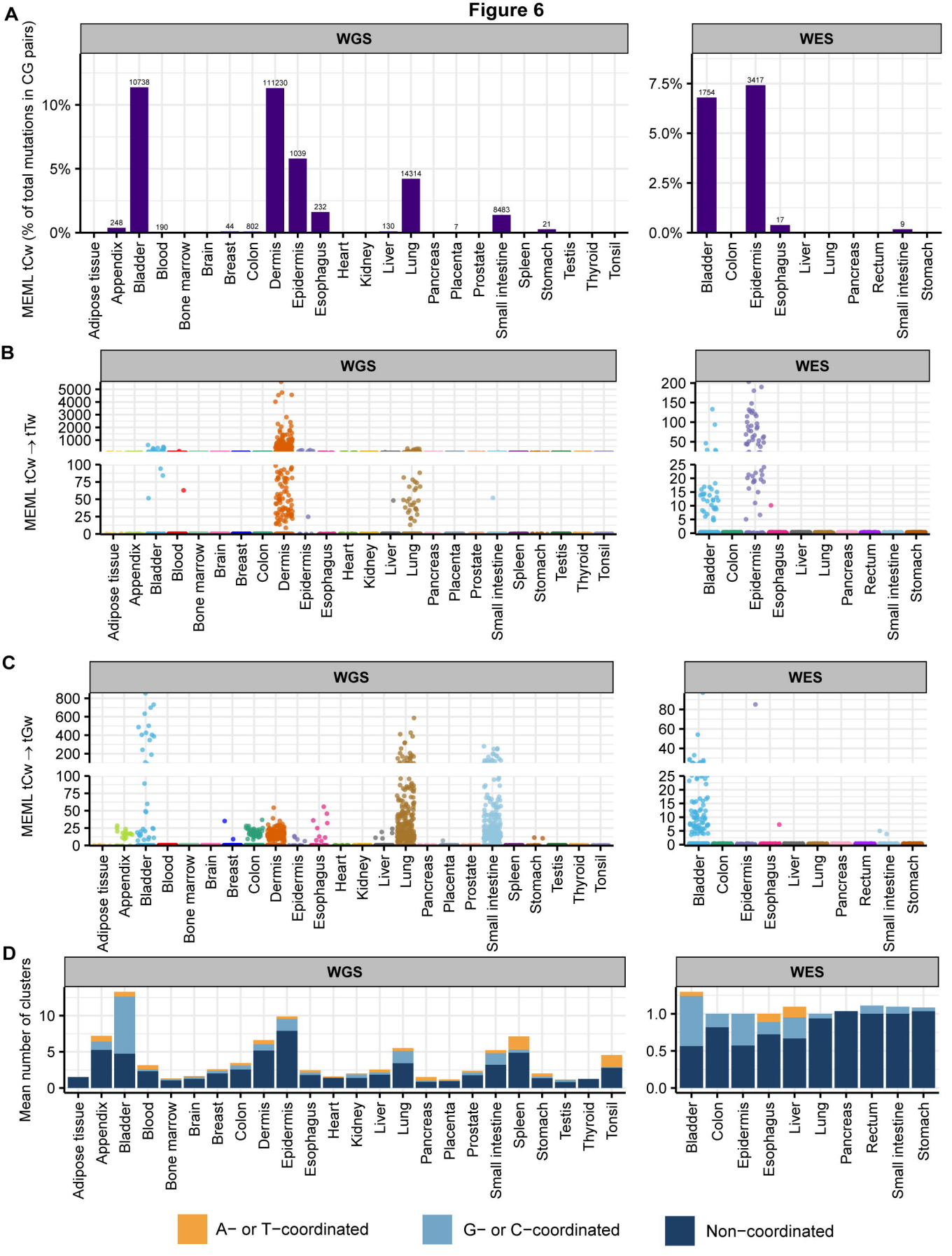

A
WGS
WES
MEML tCw (% of total mutations in CG pairs)
10%
5%
0%
7.5%
5.0%
2.5%
0.0%
248
10738
190
44
802
111230
1039
232
130
14314
7
8483
21
1754
3417
17
9
Adipose tissue
Appendix
Bladder
Blood
Bone marrow
Brain
Breast
Colon
Dermis
Epidermis
Esophagus
Heart
Kidney
Liver
Lung
Pancreas
Placenta
Prostate
Small intestine
Spleen
Stomach
Testis
Thyroid
Tonsil
Rectum
B
MEML tCw → tTw
C
MEML tCw → tGw
D
Mean number of clusters
A− or T−coordinated
G− or C−coordinated
Non−coordinated

Figure 7

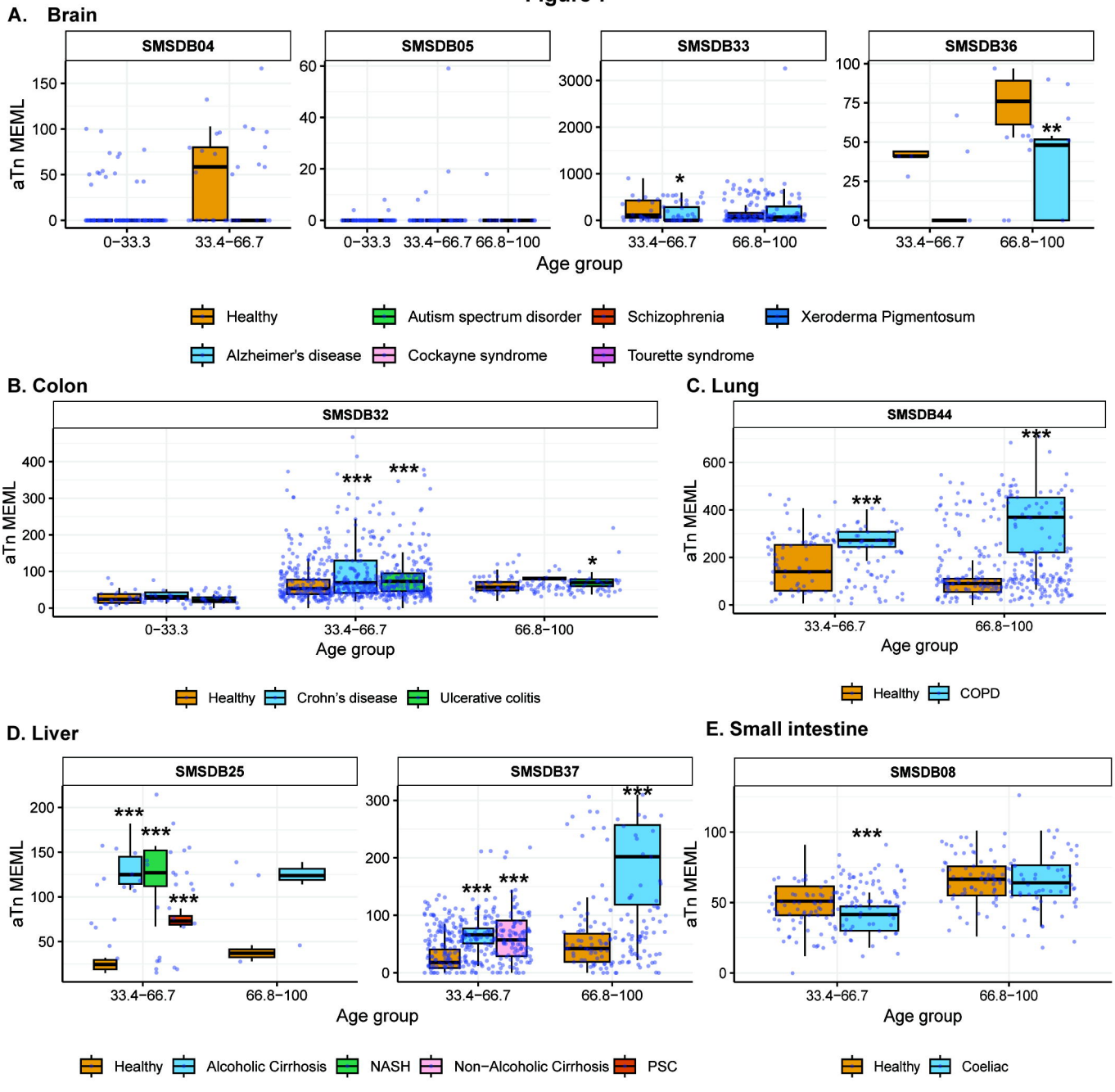

A. Brain
SMSDB04
SMSDB05
SMSDB33
SMSDB36
aTn MEML
0–33.3
33.4–66.7
66.8–100
Age group
Healthy
Autism spectrum disorder
Schizophrenia
Xeroderma Pigmentosum
Alzheimer's disease
Cockayne syndrome
Tourette syndrome
B. Colon
SMSDB32
Healthy
Crohn's disease
Ulcerative colitis
C. Lung
SMSDB44
COPD
D. Liver
SMSDB25
SMSDB37
Alcoholic Cirrhosis
NASH
Non–Alcoholic Cirrhosis
PSC
E. Small intestine
SMSDB08
Coeliac

Figure 8

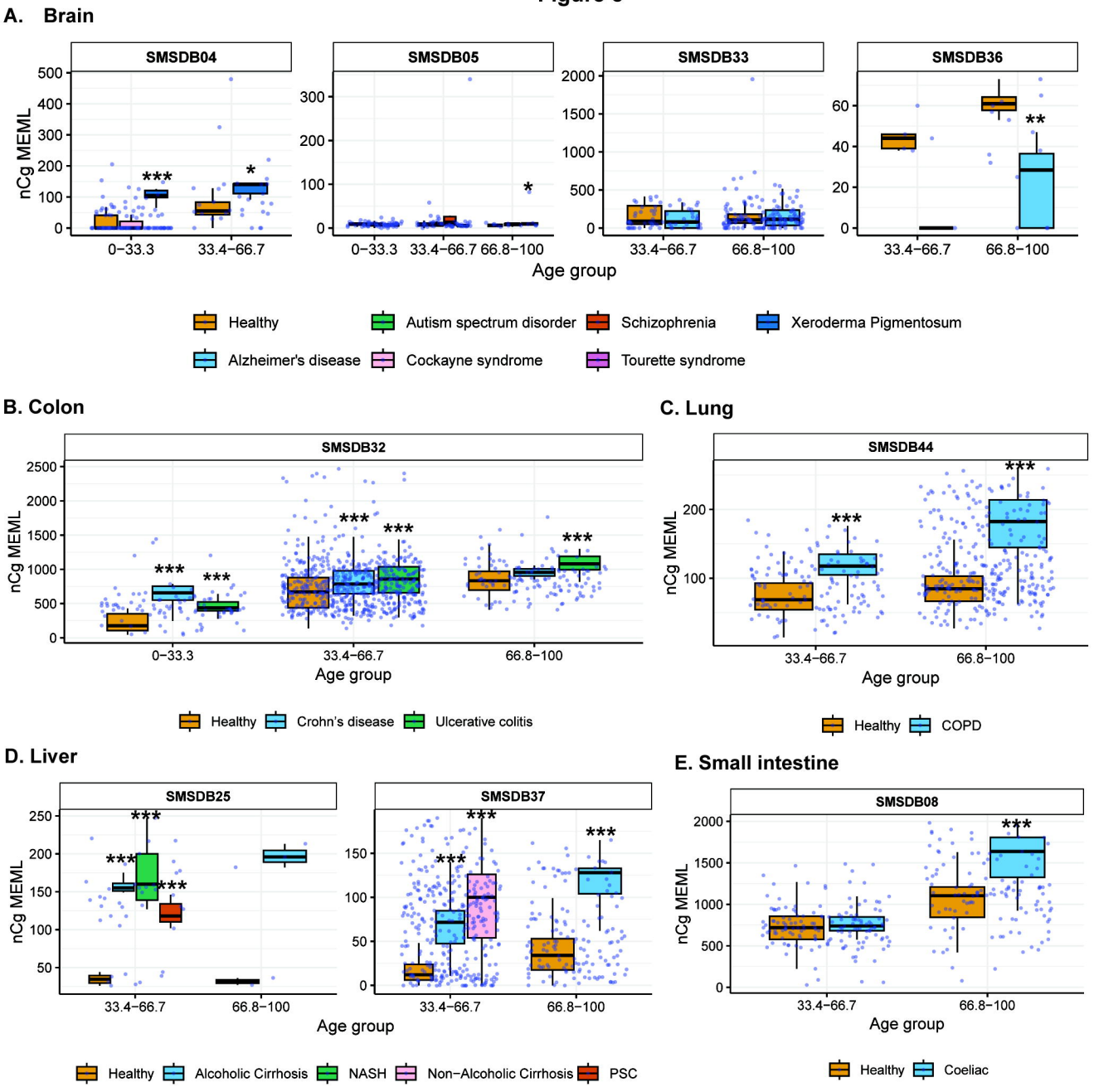
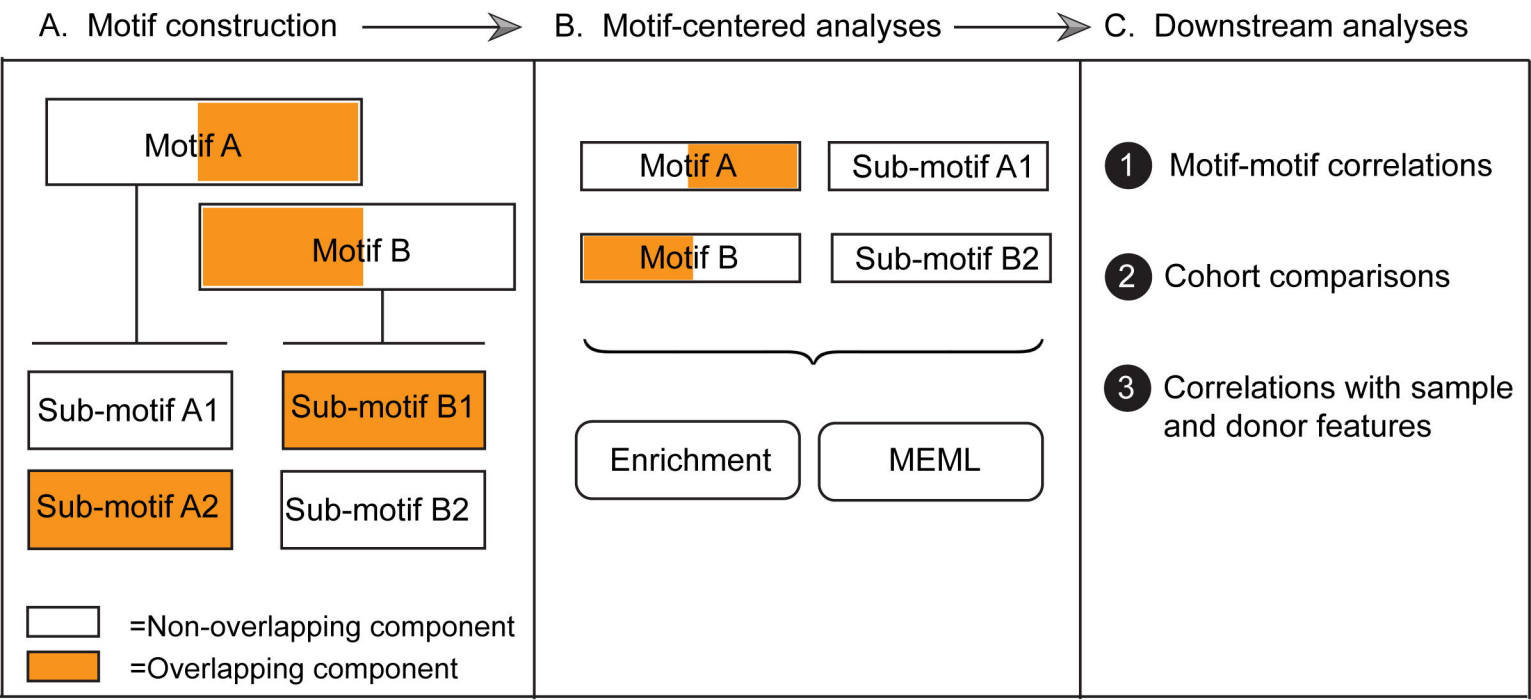

A. Brain
SMSDB04
SMSDB05
SMSDB33
SMSDB36
nCg MEML
0–33.3
33.4–66.7
66.8–100
Age group
Healthy
Autism spectrum disorder
Schizophrenia
Xeroderma Pigmentosum
Alzheimer's disease
Cockayne syndrome
Tourette syndrome
B. Colon
SMSDB32
Crohn's disease
Ulcerative colitis
C. Lung
SMSDB44
COPD
D. Liver
SMSDB25
SMSDB37
Alcoholic Cirrhosis
NASH
Non–Alcoholic Cirrhosis
PSC
E. Small intestine
SMSDB08
Coeliac

Figure 9


A. Motif construction
B. Motif-centered analyses
C. Downstream analyses
Motif A
Motif B
Sub-motif A1
Sub-motif B1
Sub-motif A2
Sub-motif B2
=Non-overlapping component
=Overlapping component
Motif A
Sub-motif A1
Motif B
Sub-motif B2
Enrichment
MEML
1 Motif-motif correlations
2 Cohort comparisons
3 Correlations with sample and donor features