Correlation of a pan-tissue epigenetic mitotic clock with tumor cell-of-origin

Tianyu Zhu [1], Huige Tong [1], Zhaozhen Du [1] and Andrew E. Teschendorff [1,*]

1. CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institute for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China.

* Corresponding author: Andrew E. Teschendorff andrew@sinh.ac.cn

Abstract

The cumulative number of stem cell divisions in a tissue, also known as mitotic age, is thought to be a major determinant of cancer-risk. Somatic mutational and DNA methylation (DNAm) clocks are promising tools to molecularly track mitotic age, yet their relationship is unexplored and their potential for cancer risk prediction in the tumor cell-of-origin remains to be demonstrated. Using advanced cell-type deconvolution methods in conjunction with a novel pan-tissue epigenetic mitotic clock called stemTOC, we here demonstrate that a sample's mitotic age correlates with its putative tumor cell-of-origin fraction. We show that major cancer risk factors, including age, smoking and obesity-associated inflammation, increase the mitotic age of relevant normal tissues. StemTOC correlates with a somatic mutational clock-like signature, yet data indicates that the former is a more sensitive proxy for mitotic age. These results support the view that DNAm can track mitotic age in the tumor cell of origin of normal, preneoplastic and cancer tissues. StemTOC is freely available and could be adapted for future cancer-risk prediction strategies.

Keywords: DNA methylation; epigenetic clock; cancer; somatic mutations; stem-cell; ageing; mitotic

There is growing evidence that the risk of neoplastic transformation of any given tissue in any given individual is influenced by the cumulative number of stem-cell divisions (aka mitotic age), which the underlying cell-of-origin has undergone [1-3]. This mitotic age is not only determined by the intrinsic turnover rate of the tissue ("the bad-luck" hypothesis), but also by exogenous cancer risk factors (e.g. smoking, chronic inflammation) that modulate this baseline stem-cell division rate [4-6 7 8-12]. Underpinning this association of mitotic age with cancer risk is the gradual accumulation of molecular alterations following stem-cell division, eventually predisposing specific subclones to neoplastic transformation [5,13-18]. These molecular alterations, if measured, could be used to estimate mitotic age [17,19-28].

Among the potential molecular alterations that could be used to measure mitotic age, somatic mutations and DNA methylation (DNAm) have emerged as the most promising ones [17,20,21,29]. DNAm-based mitotic-like clocks hypothesized to track the cumulative number of DNA methylation errors arising during cell-division have been proposed [22-24,30-32] and shown to be universally accelerated in cancer and precancerous lesions compared to age-matched normal-tissue, yet confounders such as cell-type heterogeneity (CTH) have cast doubt on the biological and clinical significance of these findings. Moreover, the relation between DNAm-clocks and clock-like mutational

signatures remains uncharted territory.

Here we build and validate a pan-tissue epigenetic mitotic clock, called stemTOC, and combine it with advanced cell-type deconvolution methods [33,34] to demonstrate that the mitotic age of a sample correlates with the fraction of its putative tumor cell-of-origin, thus establishing a direct link between mitotic age and tumor progression. By cross-correlating stemTOC to proposed clock-like mutational signatures, we establish that the number of C>T mutations representing deamination of 5-methylcytosine, is a marker of mitotic age. We further show that normal tissues exposed to cancer risk factors, or at risk of cancer development, display an increased mitotic age relative to age-matched normal tissue. Collectively, this data highlight the advantage of DNAm changes over clock-like mutational signatures [17,35] as a means of tracking mitotic age in normal tissues at cancer-risk, whilst also adding substantial evidence to the hypothesis put forward by Tomasetti and Vogelstein [1], that the mitotic age of a tissue is a major cancer-risk factor.

Results

Construction of stemTOC

Compared to previous mitotic clocks, we here aimed to build a novel mitotic clock that is substantially less confounded by CTH and chronological age. To reduce confounding by CTH it is critical to build a mitotic clock with CpGs that are not cell-type specific in an appropriate ground-state [36]. To this end, we collected Illumina 450k/EPIC DNAm profiles from 86 fetal samples encompassing 13 fetal tissue-types (**Methods**), to select promoter-associated CpGs that are constitutively unmethylated across all fetal tissue-types (**Fig.1a**). Under this model, cell-type specific DNAm differences that may arise at these CpGs in aged tissues is due to differences in the cell-type specific mitotic rates. To avoid confounding by chronological age, we also required that CpGs undergo significant DNA hypermethylation with increased population-doublings (PDs) *in-vitro* across a range of different normal cell-lines, and that they simultaneously do not undergo hypermethylation in these same cell-lines when treated with a cell-cycle inhibitor (e.g mitomycin) or under reduced growth-promoting conditions (i.e. serum deprivation) (**Methods**) [30]. These requirements resulted in 629 mitotic CpG candidates (denoted "vitro-mitCpGs", **Fig.1a**). In order to avoid confounding by cell-culture effects, we next required these CpGs to also undergo significant DNA hypermethylation with chronological age in three separate large *in-vivo* blood DNAm datasets (**Fig.1b**). Since the cell-line data allows removal of most CpGs that accumulate DNA hypermethylation purely because of "passage of time" (and hence also "chronological age"), this additional requirement is designed to ascertain that these vitro-mitCpGs do

display DNA hypermethylation *in-vivo,* in line with the fact that mitotic age increases with chronological age. Blood-tissue was chosen for 3 reasons. First, blood has a relatively high turnover rate and it is only for such tissues that we can expect a correlation between chronological and mitotic age. Second, the availability of many large whole blood DNAm datasets ensures adequate power to detect DNAm-changes associated with chronological and mitotic age. Third, for blood-tissue we can adjust for CTH at high resolution (12 immune cell-subtypes) [37], which further ensures that the observed DNAm changes are not due to shifts in underlying cell-type proportions [38]. At the end of this step, we thus obtained a reduced subset of 371 mitotic CpG candidates, that we call "vivo-mitCpGs" (**Fig.1b**). Using eFORGE2 [39,40], we observed strong and exclusive enrichment for DNase Hypersensitive Sites (DHSs) as defined in hESCs (**SI fig.S1**). Among chromatin states, we observed strong enrichment for bivalent TSS, which was strongest for those also defined in hESCs (**SI fig.S2**). Correspondingly, we observed strong enrichment for both H3K27me3 and H3K4me3 marks, although notably stronger for the repressive H3K27me3 mark (**SI fig.S1**), consistent with previous observations that H3K4me3 is moderately protective of age-associated DNAm accrual [36,41].

Finally, based on the 371 vivo-mitCpGs, we derive a relative estimate of mitotic age that explicitly accounts for the underlying stochasticity of mitotic DNAm changes in normal tissues [25,38,42] (**Methods**, **Fig.1c**). This stochasticity implies that only a relatively small fraction of the vivo-mitCpGs may be tracking mitotic age in any given subject and that this specific subset may vary randomly between subjects. We build a simple yet realistic simulation model (**Methods**) to demonstrate how taking a specific upper quantile of the DNAm distribution over mitotic CpGs should yield an improved estimator of mitotic age compared to taking an average DNAm over these same CpGs (**Fig.1c**) [25,42]. To identify a sensible upper-quantile threshold, we computed upper-quantiles for the 371 vivo-mitCpGs in a real DNAm dataset encompassing 42 normal breast samples adjacent to breast cancer ("normal-tissue at risk") and 50 age-matched normal breast samples from healthy women ("normal-healthy") [42] (**Methods, Fig.1d**). This revealed substantially larger effect sizes for higher upper-quantile values consistent with DNAm changes in the normal-tissue at risk being highly stochastic. We decided on a 95% upper quantile threshold to maximize effect size without compromising sampling variability (**Fig.1d**).

Validation and benchmarking of stemTOC

One way to assess if stemTOC is measuring mitotic age in *in-vivo* human tissue samples is by cross-comparing correlations of mitotic age estimates with chronological age across tissues characterized by highly variable turnover rates, because for tissues with

a high turnover-rate the total number of stem-cell divisions will be strongly determined by the subject's chronological age, whilst for tissues with a low turnover rate, or for tissues where the mitotic age is more strongly influenced by factors operating on shorter time scales (e.g. hormonal factors), the correlation between mitotic age and chronological age should be much weaker or non-evident. Applying stemTOC to the normal-adjacent tissue data from the TCGA confirmed this: high and medium turnover tissues like colon and rectum displayed strong correlations between mitotic and chronological age, whilst hormone-sensitive tissues like breast and endometrium did not, despite these being appropriately powered (**Fig.2a-b, SI fig.S3, SI table.S1**). Similar results were obtained when estimating mitotic age in the enhanced GTEX (eGTEX) DNAm dataset, encompassing 987 samples from 9 normal tissue-types (**SI fig.S4, SI table.S2**). For instance, colon displayed a clear correlation, whilst ovary and breast did not (**SI fig.S4**). Applying stemTOC to purified immune cell populations from healthy individuals (**Methods**) also revealed relatively strong correlations with chronological age (**Fig.2c-d, SI table.S3**), in line with the hematopoietic system turning over at a relatively high rate. Correspondingly, in sorted neurons, a predominantly post-mitotic population, the correlation was not significant (**Fig.2e**). To further validate stemTOC, we compared the estimated mitotic ages of high, medium and low-turnover tissues [22]. Using the normal-adjacent samples from the TCGA encompassing 15 tissue-types, we observed a strong correlation (PCC=0.94) between stemTOC's median mitotic age of the tissue and the tissue-specific intrinsic annual rate of stem-cell division, as estimated by Tomasetti & Vogelstein [1,5] (**Fig.2f**, **Methods**). This association was confirmed by plotting stemTOC against the estimated total number of stem-cell divisions for each normal-adjacent sample on a log-scale, revealing clear differences in mitotic age between low, medium and high turnover tissues (**Fig.2g**). Benchmarking stemTOC to 6 other DNAm-based mitotic clocks (epiTOC2 [43] , epiTOC1 [22] , HypoClock [24], RepliTali [30] and EpiCMIT-hyper/hypo [44]), we observed that clocks based on hypermethylated CpGs (i.e. stemTOC, epiTOC1/2 and epiCMIT-hyper) displayed stronger correlations with age than clocks based on hypomethylated CpGs (HypoClock, RepliTali, epiCMIT-hypo) (**Fig.2b**). Similar patterns were observed in the normal-tissue data from eGTEX (**SI fig.S4**) as well as in the sorted immune-cell subsets (**Fig.2d**). StemTOC and the hypermethylation-based clocks also displayed much stronger correlations with the Vogelstein-Tomasetti stem-cell division rates, as assessed over 15 tissue-types (**Fig.2h**), suggesting that these clocks yield more accurate mitotic age estimates during normal/healthy aging compared to those based on hypomethylation.

Correlation of stemTOC's mitotic age with tumor cell-of-origin fraction

A tumor emerges by an increased and unchecked proliferative clone that derives from

the cell-of-origin. Whilst the cell-of-origin may be dedifferentiated or may have undergone trans-differentiation before the emergence of invasive cancer, we reasoned that tumors with a higher proportion of the cell-of-origin would display a higher mitotic age if mitotic age is a key marker of tumor progression. To estimate the cell-of-origin fraction in the TCGA samples we applied our recently validated EpiSCORE algorithm[33] and DNAm-atlas[34] encompassing tissue-specific DNAm reference matrices for 13 tissue-types. Before estimating cell-type fractions however, we sought additional validation of our DNAm-atlas using an independent whole-genome bisulfite-sequencing (WGBS) DNAm-atlas of 102 sorted samples from Loyfer et al [45] encompassing 15 tissue-subtypes and 17 cell-types (**Methods**). Applying EpiSCORE with the relevant tissue-specific DNAm reference matrices from our atlas, we were able to correctly predict the cell-type in Loyfer's WGBS DNAm-atlas with an overall accuracy of 85% (**Fig.3a**). Given this good validation, we next applied EpiSCORE to the TCGA samples (**SI tables.S4-S20**), which revealed strong correlations between mitotic age and tumor cell-of-origin fraction, specially for those tumor types where the cell-of-origin is reasonably well established (**Fig.3b**). For instance, colon and rectal adenocarcinoma (COAD & READ) displayed very strong correlations between stemTOC's mitotic age and the epithelial cell-fraction (**Fig.3b**). Mitotic age in luminal and basal breast cancer displayed strongest correlations with the luminal and basal fractions, respectively. In the case of pancreatic adenocarcinoma (PAAD), mitotic age was most strongly correlated with the ductal fraction. Liver hepatocellular- (LIHC) and cholangio- (CHOL) carcinoma displayed strongest correlations with hepatocyte and cholangiocyte fractions, as required. Mitotic age in skin cutaneous melanoma (SKCM) displayed the strongest correlations with the melanocyte fraction. Lung squamous cell (LUSC) and lung-adeno (LUAD) carcinoma displayed strongest correlations with basal and alveolar epithelial fractions, respectively. It is worth stressing that for each tumor-type, the strongest correlation with mitotic age was attained by the presumed tumor cell-of-origin **(SI fig.S5a**). These data clearly indicate that, despite the CTH of TCGA samples, the mitotic age estimates from stemTOC are tracking DNAm changes in the cell-of-origin of the respective tumor types.

Mitotic age estimates can also inform on the potential cell-of-origin in tumor-types where this is controversial or less well-established. For instance, in low-grade glioma (LGG) and glioblastoma multiforme (GBM), mitotic age correlated most strongly with astrocyte fraction (**Fig.3c, SI fig.S5b**), favoring an astrocyte-like progenitor as a putative cell of origin [46]. Prostate adenocarcinoma (PRAD) has long been linked to a luminal phenotype [47,48], yet studies have also indicated a potential role for basal cells in the initiation of this cancer-type [49]. Our analysis is consistent with the known luminal-cell expansion associated with PRAD (**Fig.3c, SI fig.S5b**), and the prevailing view that the cell-of-origin of human PRAD is a luminal progenitor cell [50-52].

Of note, the correlation strengths with cell-of-origin fraction obtained with stemTOC were significantly stronger than those obtained with the other 6 clocks (**Fig.3d-e**). Moreover, none of the other 6 mitotic clocks achieved 100% accuracy in predicting the putative cell-of-origin (**SI fig.S5c**). However, all mitotic clocks were universally accelerated in TCGA cancer-types, as compared to their respective age-matched normal-adjacent tissue (**SI fig.S6).** In summary, the correlation of stemTOC's mitotic age with the tumor cell-of-origin fraction underscores its potential to quantify cancer risk.

Correlation with cell-of-origin fraction in precancerous lesions and cancer-risk

We next assembled 9 DNAm datasets representing different normal healthy tissues and corresponding age-matched precancerous conditions (**Methods**). Mitotic age, as estimated with StemTOC, was significantly higher in the precancerous tissue in each of these datasets (**Fig.4a**). For instance, stemTOC's mitotic age could discriminate neoplasia from benign lesions in the prostate, intestinal metaplasia from normal gastric mucosa, or Barrett's esophagus from normal esophagus. Some of the other clocks did not provide the expected discrimination, and a formal comparison of all clocks across the 9 datasets revealed an overall improved performance of stemTOC (**SI fig.S7**). Using EpiSCORE we next estimated cell-type fractions in these datasets, and broadly speaking, for most tissue-types we observed correlations of the mitotic age with tumor cell-of-origin fraction in the precancerous lesions and in cancer itself, but not in the histologically normal tissue, although in many cases the number of samples with normal histology was much lower, which limits power to detect associations in this subgroup (**SI fig.S8**). For instance, in colon adenoma there was a clear correlation (R=0.75, P=3e-8, n=39) but not so in normal colon (R=0.44, P=0.27, n=8), likely due to the much lower number of normal samples (**SI fig.S8**). In premalignant cholangiocarcinoma (CCA) lesions, there was a strong correlation (R=0.45, P<0.001, n=60), which was not observed in normal liver samples (R=0.16, P=0.25, n=50), probably because the cholangiocyte fraction is much lower in these normal samples (**SI fig.S8**). Correlations between stemTOC's mitotic age with tumor cell-of-origin fraction in normal-adjacent tissue from the TCGA was broadly speaking also observed (e.g. COAD, BRCA, PRAD), but there were also exceptions (e.g. KIRP, LUAD) (**SI fig.S9**). Of note, using an upper quantile over the 371 stemTOC CpGs to define the mitotic age generally resulted in larger effect sizes as well as stronger correlations with cancer-status and tumor cell of origin fraction compared to taking an average, highlighting the importance of taking the underlying stochasticity of DNAm patterns into account (**SI fig.S10**). Overall, these data underscore the potential of stemTOC's mitotic age to indicate cancer risk in precancerous lesions.

Correlation of mitotic age with smoking and obesity-associated inflammation

Smoking and obesity are two main cancer risk factors that promote inflammation and which can increase the tissue's intrinsic rate of stem-cell division [7,53]. Hence, one would expect the mitotic age of a tissue to correlate with the level of exposure to such cancer-risk factors. To assess this, we analyzed an Illumina 450k DNAm data set of 790 buccal swabs from healthy women all aged 53 at sample draw (**Methods**) [54]. Buccal swabs contain approximately 50% squamous epithelial and 50% immune cells [55] and we reasoned that the mitotic age in these buccal swabs should correlate with both the squamous epithelial fraction as well as lifelong smoking habit. Estimating epithelial and immune-cell fractions following our validated HEpiDISH procedure [55], we observed a strong correlation of stemTOC's mitotic age with the squamous epithelial fraction, but importantly, also a significant positive correlation with smoking-status, which was independent of epithelial fraction (Multivariate linear regression, P=5e-5, **Fig.4b**). Epithelial cells in lung-tissue serve as the cell-of-origin of lung-cancer, and so we next estimated stemTOC's mitotic age in over 200 normal lung-tissue samples from eGTEX [56], also encompassing smokers, ex-smokers and never-smokers. Fractions for 7 cell-types including epithelial cells were estimated using EpiSCORE (**Methods, Fig.4c**) [34]. In this cohort, donors were from different ages, and correspondingly, we observed a strong correlation between stemTOC's mitotic and chronological age (**Fig.4c**). Importantly, multivariate regression analysis including age, epithelial fraction and smoking-status, revealed a significant positive correlation of mitotic age with smoking-exposure, with the mitotic age of older individuals displaying bigger differences between smokers and non-smokers (**Fig.4c**).

As another example, we focused on an EPIC DNAm dataset of liver-tissue samples from 325 obese individuals diagnosed with non-alcoholic fatty liver disease (NAFLD) [57], of which 210 displayed no fibrosis (grade-0), with the rest displaying severe fibrosis (grade 3 to 4) (**Methods**). Fractions for 5 cell-types including hepatocytes were estimated using EpiSCORE (**Methods**) [34], which confirmed the known reduction of hepatocyte fraction in NAFLD [57] (**Fig.4d**). We observed a strong correlation of stemTOC's mitotic age with chronological age, as well as with disease stage, both of which were significant in a multivariate regression analysis that also included hepatocyte fraction (**Fig.4d**). Thus, the increased mitotic age with NAFLD-stage is observed despite the reduction in hepatocyte fraction, attesting to stemTOC's sensitivity. Overall, these data support the view that stemTOC can track increased mitotic age in disease-relevant tissues exposed to major cancer-risk factors.

stemTOC correlates with a clock-like somatic mutational signature

Next, we explored the relationship of stemTOC with somatic mutational "clock-like"

signatures [17,35,58]. Among the somatic mutational signatures derived from the TCGA cancer-samples, two single-base substitution signature SBS1 and SBS5 (here termed MS1 and MS5) have been shown to be associated with chronological age, with MS1 in particular being linked to deamination of methylated CpG dinucleotides resulting in C>T mutations [17]. Given that somatic mutations in normal tissue specimens with matched DNAm data have not yet been extensively characterized, we asked if the mutational loads in TCGA cancer samples correlates with the estimated total (lifetime) number of stem-cell divisions (TNSC), as derived using the Vogelstein-Tomasetti estimates and chronological age at diagnosis (**Methods**). As remarked previously [58], although cancers display clear acceleration of mitotic age, a broad correlation between a tumor's mutational signature load and the baseline (normal) number of stem-cell divisions is expected if the signature is of a mitotic nature. Vindicating Alexandrov et al [17], MS1 displayed a correlation with the total number of stem-cell divisions (TNSC) across cancer-types, whilst MS5 did not (**Fig.5a-b**). To check that this correlation is independent of chronological age, we computed for each cancer-type the median mutational load per Mbp and calendar year over all samples of a given cancer-type, and compared it to the intrinsic rate of stem-cell division of the corresponding normal tissue-type, revealing a positive correlation for MS1 but not for MS5 (**Fig.5c-d**). Like MS1, stemTOC displayed a clear positive correlation with TNSC across tumors (**Fig.5e**), which was also independent of age (**Fig.5e-f**). However, unlike MS1, stemTOC displayed a saturation effect, with highly proliferative cancer-types displaying high stemTOC-values despite the relatively low turnover rate of the underlying normal tissue (e.g. lung cancers). Given that in normal-adjacent tissues, stemTOC displays correlations with TNSC without evidence of a saturation effect (**Fig.2e**), this suggests that stemTOC is a more sensitive tracker of mitotic age than somatic mutations.

Next, we expanded the analysis to 27 TCGA cancer-types, correlating the median stemTOC's mitotic age of each cancer-type to the corresponding median MS1 and MS5 loads. This revealed a significant correlation with MS1, but not for MS5, further attesting to the more mitotic-like nature of MS1 (**Fig.5g-h**). Of note, stemTOC correlated with MS1 within most TCGA cancer-types, even after adjusting for chronological age (**SI fig.S11**). However, in some cancer types (e.g. lung squamous cell carcinoma) no correlation was evident (**SI fig.S11**). This indicates that although stemTOC and MS1 both approximate mitotic age, that they are also distinct. In line with this, the strong correlation of stemTOC's mitotic age with tumor cell of origin fraction was not seen if one were to use the MS1-load as a proxy for mitotic age (**SI fig.S12**). Interestingly, correlations of MS1-load with CPE-based tumor purity estimates [59] were weaker than with EpiSCORE-derived tumor cell-of-origin fractions (**SI fig.S13**). Overall, these data indicate broad agreement between DNAm and somatic

mutational mitotic clocks, but with DNAm being a more sensitive marker of mitotic age.

Discussion

Despite previous studies establishing a correlation between mitotic age and cancer-risk [1,2,4,5], explicit demonstration using molecular data from human samples has been lacking. By applying a pan-tissue epigenetic mitotic clock (stemTOC) in conjunction with a state-of-the-art cell-type deconvolution method, we have here established a concrete direct link between the mitotic age of a sample and its tumor cell-of-origin fraction in each of 15 cancer-types. stemTOC also displayed accelerations in precancerous lesions from 8 normal tissue-types, as well as with smoking exposure and obesity in oral/lung and liver tissues, respectively. Collectively, these findings add substantial weight to the hypothesis put forward by Tomasetti and Vogelstein [1] , that the mitotic age of the cell-of-origin is a major determinant of cancer-risk, and that it increases with exposure to exogenous cancer-risk factors.

It is illuminating to discuss these results in relation to what has been found using somatic mutational signatures [17,35]. Many somatic mutational signatures were derived from TCGA cancer samples, of which two (MS1 and MS5) displayed frequent correlations with chronological age across cancer samples [17], and with MS1 hypothesized to be associated with cell-proliferation. By cross-comparing the reported mutational loads of TCGA cancer-samples with their estimated total number of stem-cell divisions, we have here confirmed that MS1 is mitotic-like and that MS5 is not. In line with this, stemTOC displayed a significant but moderate correlation with MS1 and not with MS5. Although recent studies have demonstrated a correlation of MS1 with chronological age in a number of normal healthy tissue-types [60,61], the analogous correlation of DNAm-based mitotic ages with chronological age has been widely demonstrated across substantially more normal tissue-types, including precancerous lesions. This largely underscores the advantage of DNAm over somatic mutations as a technically more feasible and sensitive molecular substrate to track mitotic age in normal tissues.

When correlating a molecular clock's mitotic age to chronological age in cancer samples, it is worth pointing out that for highly proliferative tumor-types arising from tissues with a relatively low turnover-rate (e.g. liver, pancreas, lung), the estimated mitotic age would not be expected to correlate well with age at diagnosis since for these cancer samples the majority of cell-divisions would have occurred after tumor onset. This is exactly the pattern seen for stemTOC, which displayed a more non-linear correlation and saturation effect with age across tumor-samples. In contrast, MS1

displayed a more linear pattern. Correspondingly, stemTOC displays much stronger linear correlations with age when assessed in normal-tissues. Once again this underscores the higher sensitivity of DNAm changes (stemTOC) to track increased cell-divisions in normal tissue and tumors. This is a critically important insight for future efforts to track clonal evolution in precancerous states and on short time-scales [62].

A benchmark of stemTOC against previous DNAm-based clocks further revealed that clocks anchored on sites gaining methylation (stemTOC and epiCMIT-hyper) appear to provide better proxies of mitotic age in normal adult tissues compared to HypoClock or RepliTali, which are largely based on hypomethylation. Thus, whilst recent studies have advocated tracking mitotic age via loss of DNAm at specific CpGs mapping to late-replicating partially methylated domains [24], our data reinforces the alternative view that DNAm gains at specific genes that are initially unmethylated in fetal tissue, is an equally good, if not better, tracker of mitotic age, specially in the context of adult normal tissue turnover [43]. Indeed, our data indicates that *in-vivo* loss of DNAm at solo-CpGs may only apply to cells that are under high replicative stress (e.g. early development, cancer or high turnover tissues like colon) [24,43].

Benchmarking stemTOC against other DNAm-based clocks also highlighted the importance of accounting for the underlying stochastic DNAm changes in aging and precancerous lesions [25,42,63]. Indeed, by approximating mitotic age with an upper quantile statistic over the pool of mitotic CpGs (as opposed to taking an ordinary average), we can better account for the inherent inter-CpG and inter-subject stochasticity, enabling the identification of DNAm outliers that very likely reflect epigenetic mosaicism and subclonal expansions [38,42]. In line with this, these DNAm outliers led to mitotic age estimates displaying stronger correlations with cancer and tumor cell-of-origin fraction.

In summary, using a mechanistic DNAm-based mitotic clock we have demonstrated a clear correlation of a sample's mitotic age with cancer risk factors and with the sample's tumor cell-of-origin fraction, supporting the view that the cumulative number of stem-cell divisions in a human tissue is a marker of cancer risk.

Methods

Illumina DNAm datasets used in construction of stemTOC

Fetal tissue DNAm sets: We obtained and normalized Illumina 450k data from the Stem-Cell Matrix Compendium-2 (SCM2) [64], as described by us previously [22]. There were a total of 37 fetal tissue samples encompassing 10 tissue types (stomach, heart, tongue, kidney, liver, brain, thymus, spleen,lung, adrenal gland). We also obtained Illumina 450k data of 15 cord-blood samples [65] and 34 fetal-tissue samples from Slieker

et al [66] encompassing amnion, muscle, adrenal and pancreas. Both sets were normalized like the SCM2 data, i.e. by processing idat files with *minfi* [67] followed by type-2 probe bias adjustment with BMIQ [68].

Cell-line data from Endicott et al [30]: The Illumina EPICv1 cell line data was downloaded from GEO under accession number GSE197512. Raw idat files were processed with *minfi* and BMIQ normalized, resulting in a normalized DNAm data matrix defined over 843,298 probes and 182 "baseline-profiling" samples. A total of 6 cell lines were used, including AG06561, AG11182, AG11546. AG16146, AG21837, and AG21859. One cell-line (AG21839), which displayed global non-monotonic DNAm patterns with population doublings (PDs) was removed.

Whole blood DNAm datasets: We used 3 Illumina whole blood datasets. One EPIC dataset encompassing 710 samples from Han Chinese was processed and normalized as described by us previously [69]. Briefly, idat files were processed with minfi, followed by BMIQ type-2 probe adjustment and due to the presence of beadchip effects, data was further normalized with ComBat [70]. Another dataset is an Illumina 450k set of 656 samples from Hannum et al [71]. The normalization of this dataset is as described by us previously [69]. Finally, we also analysed a 450k dataset from Johansson et al [72]. This dataset was obtained from NCBI GEO website under accession number GSE87571 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87571). The file "GSE87571_RAW.tar" containing the IDAT files was downloaded and processed with *minfi* R-package. Probes with P-values < 0.05 across all samples were kept. Filtered data was subsequently normalized with BMIQ, resulting in a normalized data matrix for 475,069 probes across 732 samples.

Illumina DNAm datasets from the TCGA

The SeSAMe [73] processed Illumina 450k beta value matrices were downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) with TCGAbiolinks [74]. We analyzed 32 cancer types (ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, PAAD, PCGP, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM). We did further processing of the downloaded matrices as follows: For each cancer type, probes with missing values in more than 30% samples were removed. The missing values were then imputed with impute.knn (k=5). Then, the type-2 probe bias was adjusted with BMIQ. Technical replicates were removed by retaining the sample with highest CpG coverage. Clinical information of TCGA samples was downloaded from Liu et al [75].

Illumina DNAm dataset from eGTEX

The *ChAMP* [76] processed Illumina EPIC beta valued data matrix was downloaded from GEO (GSE213478) [56]. The dataset includes 754,119 probes and 987 samples from 9 normal tissue-types (breast mammary tissue, colon transverse, kidney cortex, lung, muscle skeletal, ovary, prostate, testis and whole blood) derived from 424 GTEX subjects.

Illumina DNAm datasets of sorted immune-cell types

We analyzed Illumina 450k DNAm data from BLUEPRINT[77], encompassing 139 monocyte, 139 CD4+ T-cell and 139 neutrophil samples from 139 subjects. This dataset was processed as described by us previously [78]. In addition, we analyzed Illumina 450k DNAm data from GEO (GSE56046) encompassing 1202 monocyte and 214 naïve CD4+ T-cell samples from Reynolds et al [79], and GEO (GSE56581) encompassing 98 CD8+ T-cell samples from Tserel et al [80]. Data was normalized as described by us previously [81]. We also analyzed Illumina 450k DNAm data from EGA (EGA: EGAS00001001598) encompassing 100 B cell, 98 T cell, and 104 monocyte samples from Paul et al [82]. Data was normalized as described by us previously [82].

Illumina DNAm datasets of sorted neurons

We analyzed a total of 4 Illumina DNAm datasets of sorted neurons (neuronal nuclei NeuN+), in all cases only using normal control samples. In all cases, raw Illumina EPIC/450k DNAm files were downloaded from GEO under accession numbers GSE112179 (n=28, post-mortem frontal cortex) [83] , GSE41826 (n=29, post-mortem frontal cortex) [84], GSE66351 (n=16, post-mortem human brains) [85] and GSE98203 (n=29, post-mortem human brains) [86]. In all cases, probes not detected above background were assigned NA. CpGs with coverage>0.99 were kept (in the 2nd and 4th sets, this threshold was relaxed slightly to 0.98). Remaining NAs were imputed with impute.knn (k=5). Finally, the beta-valued data matrix was normalized with BMIQ.

Illumina DNAm datasets from normal healthy and normal "at cancer-risk" tissue

1. *Lung preinvasive dataset:* This is an Illumina 450k DNAm dataset of lung-tissue samples that we have previously published [54]. We used the normalized dataset from Teschendorff et al [54] encompassing 21 normal lung and 35 age-matched lung-carcinoma in-situ (LCIS) samples, and 462,912 probes after QC. Of these 35 LCIS samples, 22 progressed to an invasive lung cancer (ILC).
2. *Breast preinvasive dataset:* This is an Illumina 450k data set of breast tissue samples from Johnson et al [87]. Raw idat files were downloaded from GEO under accession number GSE66313, and processed with *minfi*. Probes with sample coverage<0.95 (defined as fraction of samples with detected (P<0.05) P-values were discarded. The rest of unreliable values were assigned NA and imputed with knn (k=5) [88]. After

BMIQ normalization, we were left with 448,296 probes and 55 samples, encompassing 15 normal-adjacent breast tissue and 40 age-matched ductal carcinoma in situ (DCIS) samples, of which 13 were from women who later developed an invasive breast cancer (BC).

3. *Gastric metaplasia dataset:* Raw idat files were downloaded from GEO (GSE103186) [89] and processed with *minfi.* Probes with over 99% coverage were kept and missing values imputed using impute R-package using impute.knn (k=5). Subsequently, data was intra-array normalized with BMIQ, resulting in a final normalized data matrix over 482,975 CpGs and 191 samples, encompassing 61 normal gastric mucosas, 22 mild intestinal metaplasias and 108 metaplasias. Although age information was not provided, we used Horvath's clock [90] to confirm that normal and mild intestinal metaplasias were age-matched. This is justified because Horvath's clock is not a mitotic clock [43] and displays a median absolute error of +/-3 years [90].
4. *Barrett's Esophagus and adenocarcinoma dataset:* This Illumina 450k dataset [91] is freely available from GEO under accession number GSE104707. Data was normalized as described by us previously [92]. The BMIQ normalized dataset is defined over 384,999 probes and 157 samples, encompassing 52 normal squamous epithelial samples from the esophagus, 81 age-matched Barrett's Esophagus specimens and 24 esophageal adenocarcinomas.
5. *Colon adenoma dataset* [93]*:* Illumina 450k raw idat files were downloaded from ArrayExpress E-MTAB-6450 and processed with minfi. Only probes with 100% coverage were kept. Subsequent data was intra-array normalized with BMIQ, resulting in a normalized data matrix over 483,422 CpGs and 47 samples, encompassing 8 normal colon specimens and 39 age-matched colon adenomas. Although age-information was not made publicly available, we imputed them using Horvath's clock, confirming that normals and adenomas are age-matched.
6. *Cholangiocarcinoma (CCA) dataset* [94]: Raw idat files were downloaded from GEO (GSE156299). This is an EPIC dataset and was processed with minfi. Probes with >99% coverage (fraction of samples with P<0.05) were kept. NAs were imputed with impute.knn (k=5). Subsequent data was intra-array normalized with BMIQ, resulting in a normalized data matrix over 854,026 probes and 137 samples, encompassing 50 normal bile duct specimens, 60 premalignant and 27 cholangiocarcinomas. Normal bile duct, premalignant and CCA specimens were derived from the same patient. Thus, normal and premalignant samples are mostly age-matched.
7. *Prostate cancer progression dataset:* Illumina 450k raw idat files were downloaded from GEO (GSE116338) [95] and processed with minfi. CpGs with coverage>0.95 were kept. NAs were imputed with impute.knn (k=5). Subsequent data matrix was

intra-array normalized with BMIQ. Samples included benign lesions (n=10), neoplasia (n=6), primary tumors (n=14) and metastatic prostate cancer (n=6). In this dataset, benign lesions were significantly older compared to neoplasia and primary tumors, thus this dataset provides a particularly good test that mitotic clocks are measuring mitotic age as opposed to chronological age.

8. *Oral squamous cell carcinoma (OSCC) progression dataset:* The processed beta-valued Illumina 450k DNAm data matrix was downloaded from GEO (GSE123781) [96]. Probes not detected above background had been assigned NA. CpGs with coverage>0.95 were kept. Remaining NAs were imputed with impute.knn (k=5). Then the beta matrix was BMIQ normalized. Samples included 18 healthy controls, 8 lichen planus (putative premalignant condition) and 15 OSCCs. This is the only dataset where premalignant samples older than the healthy controls.
9. *Colon adenoma dataset 2.* Processed beta-valued Illumina 450k DNAm data matrix was downloaded from GEO (GSE48684) [97], encompassing 17 normal-healthy samples, 42 age-matched colon adenoma samples and 64 colon cancer samples. Probes not detected above background were assigned NA. CpGs with coverage>0.95 were kept. Remaining NAs were imputed with impute.knn (k=5). Then the beta matrix was BMIQ normalized. Although age information was not made publicly available, ages were imputed with Horvath's clock confirming that normals and adenomas are age-matched.
10. *Normal breast Erlangen dataset:* This Illumina 450k dataset is freely available from GEO under accession number GSE69914. Data was normalized as described by us previously [42]. The BMIQ normalized dataset is defined over 485,512 probes and 397 samples, encompassing 50 normal-breast samples from healthy women, 42 age-matched normal-adjacent samples and 305 invasive breast cancers. We note that this dataset was excluded from the formal comparison of stemTOC to all other clocks, because this dataset was used to select the upper-quantile threshold in the definition of stemTOC (see later).

Illumina DNAm datasets of normal tissues exposed to cancer risk factors

1. *Buccal swabs+smoking (n=790):* This Illumina 450k DNAm dataset was generated and analyzed previously by us [54]. We used the same normalized DNAm dataset as in this previous publication. Briefly, the samples derive from women all aged 53 years at sample draw, and belong to the MRC1946 birth cohort. This cohort has well-annotated epidemiological information, including smoking-status information. Among the 790 women, 258 were never-smokers, 365 ex-smokers and 167 current smokers at sample-draw.
2. *Lung-tissue+smoking (n=204):* This Illumina EPIC DNAm dataset of normal lung tissue samples derives from eGTEX [56] and was processed as the other tissue datasets

from eGTEX. Age-information was only provided in age-groups. Smoking status distribution was 89 current smokers, 54 ex-smokers and 61 never-smokers.

3. *Liver-tissue+obesity (n=325):* This Illumina EPIC DNAm dataset of liver tissue is derived from GEO: GSE180474, and encompasses liver tissue samples from obese individuals (minimum BMI=32.6), all diagnosed with non-alcoholic fatty liver disease (NAFLD) [57]. Of the 325 samples, 210 had no evidence of fibrosis (grade-0), 55 had grade-3 fibrosis, 36 intermediate grade3-4 fibrosis and 24 grade-4 fibrosis. We downloaded the processed beta and P-values and only kept probes with 100% coverage across all samples. Data was further adjusted for type-2 probe bias using BMIQ.

Construction of stemTOC: identification of mitotic CpGs

The construction of stemTOC initially involves a careful selection of CpGs that track mitotic age. Initially, we follow the procedure as described for the epiTOC+epiTOC2 mitotic clocks, identifying CpGs mapping to within 200bp of transcription start sites and that are constitutively unmethylated across 86 fetal tissue samples encompassing 15 tissue-types. Here, by constitutive unmethylation we mean a DNAm beta-value < 0.2 in each of 86 fetal tissue samples encompassing 13 tissue-types. Of note, this means that all these CpGs occupy the same methylation state in the fetal ground-state, i.e. these CpGs are not cell-type or tissue-specific in this ground state. Next, we use the cell-line data from Endicott et al [30] to further identify the subset of CpGs that display significant hypermethylation with population doublings, but which do not display such hypermethylation when the cell-lines are treated with a cell-cycle inhibitor (mitomycin-C) or when the cell-culture is deprived from growth-promoting serum. In detail, for each of 6 cell-lines (AG06561, AG11182, AG11546, AG16146, AG21837, AD21859) we ran linear regressions of DNAm against population doublings (PD) (a total of 182 samples), identifying for each cell-line CpGs where DNAm increases with PDs (q-value (FDR) < 0.05). Out of a total of 843298 CpGs, 14,255 CpGs displayed significant hypermethylation with PDs in each of the 6 cell-lines. Of these 14,225 CpGs we next removed those still displaying hypermethylation (unadjusted P<0.05) with days in culture in cell-lines treated with mitomycin-C or in cell-lines deprived of serum. This resulted in a much reduced set of 629 “vitro-mitotic” CpG candidates. In the next step, we asked how many of these “vitro-mitotic” CpGs display age-associated hypermethylation in-vivo, using 3 separate large whole blood cohorts. The rationale for using whole blood is that this is a high turnover-rate tissue, and so chronological age should be correlated with mitotic age. Moreover, large whole blood cohorts guarantee adequate power to detect age-associated DNAm changes. And thirdly, by intersecting CpGs undergoing hypermethylation with PDs in-vitro with those undergoing

hypermethylation with age in-vivo, we are more likely to be identifying CpGs that track mitotic age in-vivo. We used the 3 large whole blood datasets as described in the previous section, each containing approximately 700 samples, whilst adjusting for all potential confounders including cell-fractions for 12 immune-cell subtypes [37]. Specifically, we used a DNAm reference matrix defined over 12 immune cell-types, that we have recently validated across over 23,000 whole blood samples from 22 cohorts [98], to estimate cell-type proportions using our EpiDISH procedure [99]. These proportions were then included as covariates when identifying age-DMCs in each cohort separately. To arrive at a final set of age-DMCs, we used the directional Stouffer method over the 3 large cohorts to compute an overall Stouffer z-statistic and P-value, selecting CpGs with $z > 0$ and $P<0.05$. Of the 629 "vitro-mitotic" CpGs, 371 were significant in the Stouffer meta-analysis of whole blood cohorts. This set of 371 CpGs defined our "vivo-mitotic" CpGs making up stemTOC.

Estimating relative mitotic age with stemTOC

Given an arbitrary sample with a DNAm profile defined over these 371 CpGs, we next define the mitotic age of the sample as the 95% upper quantile of DNAm values over the 371 vivo-mitotic CpGs. The justification for taking an upper quantile value, as opposed to taking an average is as follows: extensive DNAm data from previous studies indicate that DNAm changes that mark cancer-risk (and hence also mitotic age) are characterized by an underlying inter-CpG and inter-subject stochasticity [25,38,42,63]. Specifically, relevant DNAm changes in normal-tissues at cancer risk constitute mild outliers in the DNAm distribution representing subclonal expansions, that occur only infrequently across independent subjects, with the specific CpG outliers displaying little overlap between subjects. Thus, for a given pool of mitotic CpG candidates (i.e. the pool of 371 vivo-mitotic CpGs derived earlier), only a small subset of these may be tracking mitotic age in any given subject at-risk of tumor development. Thus, taking an average DNAm over the 371 CpGs is not optimized to capturing the effect of subclonal expansions that track the mitotic age. To understand this, we ran a simple simulation model, with parameter choices inspired from real DNAm data [25,38,42,63], for a reduced set of 20 CpGs and 40 samples representing 4 disease stages (normal, normal at-risk, preneoplastic, cancer) with 10 samples in each stage. In the initial stage, all CpGs are unmethylated with DNAm beta-values drawn from a Beta distribution *Beta(a=10,b=90)*, where *Beta(a,b)* is defined by the probability distribution

$$p(X = x|a, b) \sim x^{a-1}(1-x)^{b-1}$$

which has a mean $E[X] = a/(a+b)$ and variance $Var[X] = ab/((a+b+1)(a+b)^2)$. For each "normal at-risk" sample, we randomly selected 1-3 CpGs (precise number was drawn from a uniform distribution), simulating DNAm gains for these CpGs by drawing DNAm-values from *Beta(a=3,b=7)*. For each preneoplastic sample,

we randomly selected 5-10 CpGs with DNAm values drawn from *Beta(a=5,b=5)*. Finally, for invasive cancer, we randomly selected 11-17 CpGs with DNAm gains drawn from *Beta(a=8,b=2)*. Under this model, discrimination of normal-at-risk samples from normal healthy is difficult if one were to average DNAm over the 20 CpGs. However, taking a 95% upper quantile (UQ) of the DNAm distribution over the 20 CpGs, one can discriminate the normal at-risk samples. In this instance, a 95% UQ over 20 CpGs corresponds to taking the maximum value over these 20 CpGs. Another way to understand this, is by first identifying CpGs that display DNAm outliers in the normal at-risk group compared to normal. This can be done using differential variance statistics [25,100] to identify hypervariable CpGs, or alternatively, by finding CpGs for which a suitable upper-quantile over the normal at-risk samples is much greater than the corresponding UQ-value over the normal samples. To determine what UQ-threshold may be suitable, we analyzed our Illumina 450k DNAm dataset encompassing 50 normal-breast tissues from healthy women and 42 age-matched normal-adjacent samples ("at-risk" samples) from women with breast cancer [42]. For each of the 371 CpGs we computed the mean and UQ over the 50 normal-breast samples, and separately over the 42 normal-adjacent "at-risk" samples. We considered a range of UQ thresholds from 0.75 to 0.99. We then compared the difference (i.e. effect size) in the obtained values between the normal-adjacent and normal-healthy, averaging over the 371 CpGs. The effect size increases with higher UQ values. We selected a 95% UQ because at this threshold we maximized effect size without compromising variability (choosing higher UQs leads to increased random variation). Although in this analysis, the UQ is taking across samples, it is reasonable, given the underlying stochasticity of the patterns, to apply the same UQ-threshold across CpGs. Thus, for any independent sample, we define the relative mitotic age of stemTOC as the 95% upper-quantile of the DNAm distribution defined over the 371 CpGs. We note however that results in this manuscript are not strongly dependent on the particular UQ value, i.e results are generally very robust to UQ values in the range 75-95%.

Estimating cell-type fractions in solid tissues with EpiSCORE and HEpiDISH

In this work we estimate the proportions of all main cell-types within tissues from the TCGA using our validated EpiSCORE algorithm [33] and its associated DNAm atlas of tissue-specific DNAm reference matrices [34]. This atlas comprises DNAm reference matrices for lung (7 to 9 cell-types: alveolar epithelial, basal, other epithelial, endothelial, granulocyte, lymphocyte, macrophage, monocyte and stromal), pancreas (6 cell-types: acinar, ductal, endocrine, endothelial, immune, stellate), kidney (4 cell-types: epithelial, endothelial, fibroblasts, immune), prostate (6 cell-types: basal, luminal, endothelial, fibroblast, leukocytes, smooth muscle), breast (7 cell-types: luminal, basal,

fat, endothelial, fibroblast, macrophage, lymphocyte), olfactory epithelium (9 cell-types: mature neurons, immature neurons, basal, fibroblast, gland, macrophages, pericytes, plasma, T-cells), liver (5 cell-types: hepatocytes, cholangiocytes, endothelial, Kupffer, lymphocytes), skin (7 cell-types: differentiated and undifferentiated keratinocytes, melanocytes, endothelial, fibroblast, macrophages, T-cells), brain (6 cell-types: astrocytes, neurons, microglia, oligos, OPCs and endothelial), bladder (4 cell-types: endothelial, epithelial, fibroblast, immune), colon (5 cell-types: endothelial, epithelial, lymphocytes, myeloid and stromal) and esophagus (8 cell-types: endothelial, basal, stratified, suprabasal and upper epithelium, fibroblasts, glandular, immune). Hence, when estimating cell-type fractions in the TCGA tissue samples, we were restricted to those tissues with an available DNAm reference matrix. EpiSCORE was run on the BMIQ-normalized DNAm data from the TCGA with default parameters and 500 iterations. EpiSCORE was also used to estimate cell-type fractions in solid tissues from non-TCGA datasets. For instance, we used EpiSCORE to estimate 7 lung cell-type fractions (endothelial, epithelial, neutrophil, lymphocyte, macrophage, monocyte and stromal) in normal lung-tissue samples from eGTEX [56], and 5 liver cell-type fractions (cholangiocyte, hepatocyte, Kupffer, endothelial and lymphocyte) in liver tissue [57]. For the buccal swab dataset, because buccal swabs only contain squamous epithelial and immune-cells, we used the validated HEpiDISH algorithm and associated DNAm reference matrix to estimate these fractions in this tissue [55].

Validation of EpiSCORE fractions using a WGBS DNAm-atlas

Whilst EpiSCORE has already undergone substantial validation [34], here we decided to further validate it against the recent human WGBS DNAm-atlas from Loyfer et al [45], which comprises 207 sorted samples from various tissue-types. Briefly, we downloaded the beta-valued data and bigwig files (hg38) from the publication website. We required CpGs to be covered by at least 10 reads. We then averaged the DNAm values of CpGs mapping to within 200bp of the transcription start sites of coding genes, resulting in a DNAm data matrix defined over 25,206 gene promoters and 207 sorted samples. The 207 sorted samples represent 37 cell-types from 42 distinct tissues (anatomical regions). We validate EpiSCORE in this WGBS-atlas by asking if it could predict the corresponding cell-type. This was only done for tissues for which there is an EpiSCORE DNAm reference matrix. For instance, for breast tissue, the WGBS DNAm-atlas profiled 3 luminal and 4 basal epithelial samples. Hence, for breast tissue, we applied EpiSCORE with our breast DNAm reference matrix, derived via imputation from a breast-specific scRNA-Seq atlas, to these 7 sorted samples to assess if it can predict the basal/luminal subtypes using a maximum "probability" criterion (using the estimated fraction as a probability). This analysis was done for WGBS sorted cells from 15 anatomical sites (brain, skin, colon, breast, bladder, liver, esophagus, pancreas,

prostate, kidney tubular, kidney glomerular, lung alveolar, lung pleural, lung bronchus and lung interstitial) using our EpiSCORE DNAm reference matrices for brain, skin, colon, breast, bladder, liver, esophagus, pancreas, prostate, kidney and lung). An overall accuracy was estimated as the number of sorted WGBS samples where the corresponding cell-type was correctly predicted.

Benchmarking against other DNAm-based mitotic clocks

We benchmarked stemTOC against 6 other epigenetic mitotic clocks: epiTOC [22], epiTOC2 [43], HypoClock [24], RepliTali [30], epiCMIT-hyper and epiCMIT-hypo [44]. Briefly, epiTOC gives a mitotic score called pcgtAge, which is an average DNAm over 385 epiTOC-CpGs. In the case of epiTOC2, we used the estimated total cumulative number of stem-cell divisions (tnsc). In the case of HypoClock, the score is given by the average DNAm over 678 solo-WCGWs with representation on Illumina 450k arrays. Of note, for Hypoclock, smaller values mean a larger deviation from the methylated ground state. In the case of RepliTali, which was trained on EPIC data, the score is calculated with 87 RepliTali CpGs and their linear regression coefficients as provided by Endicott et al [30]. Of the 87 RepliTali CpGs, only 30 are present on the Illumina 450k array. As far as epiCMIT [44] is concerned, this clock is based on two separate lists of 184 hypermethylated and 1164 hypomethylated CpGs. Because the biological mechanism by which CpGs gain or lose DNAm during cell-division is distinct, the strategy recommended by Ferrer-Durante to compute an average DNAm over the two lists to then select the one displaying the biggest deviation from the ground state as a measure of mitotic age, is in our opinion not justified. Their strategy could conceivably result in the hypermethylated CpGs being used for one sample, and the hypomethylated CpGs being used for another. Instead, in this work we separately report the average DNAm of the hypermethylated and hypomethylated components. Thus, for the hypermethylated CpGs, the average DNAm over these sites defines the "epiCMIT-hyper" clock's mitotic age, whereas for the hypomethylated CpGs we take 1-average(hypomethylated CpGs) as a measure of mitotic age, thus defining the "epiCMIT-hypo" clock.

We compare all mitotic clocks using the following evaluation strategies. In the analysis correlating mitotic ages to the fractions of the tumor cell-of-origin in the various TCGA cancer-types, we compare the inferred Pearson Correlation Coefficients (PCCs) between each pair of clocks across the TCGA cancer-types using a one-tailed Wilcoxon rank sum test. Likewise, in the analysis where we correlate mitotic age to chronological age in the normal-adjacent samples from the TCGA, in the normal samples from eGTEX and the sorted immune-cell subsets, we compare the inferred PCCs between each pair of clocks across the independent datasets using a one-tailed Wilcoxon rank

sum test. Finally, when assessing the clocks for predicting cancer-risk, for each of the 9 datasets with normal-healthy and age-matched "normal at-risk" samples and for each clock, we first computed an AUC-metric, quantifying the clock's discriminatory accuracy. For each pair of mitotic clocks we then compare their AUC values across the 9 independent datasets using a one-tailed Wilcoxon rank sum test.

Comparison to stem-cell division rates and somatic mutational signatures

We obtained estimated tissue-specific intrinsic stem-cell division rates of 13 tissue-types (bladder, breast, colon, esophagus, oral, kidney, liver, lung, pancreas, prostate, rectum, thyroid and stomach) from Vogelstein & Tomasetti [1,5], supplemented with estimates for skin and blood [43], corresponding to 17 TCGA cancer-types (BLCA, BRCA, COAD, ESCA, HNSC, KIRP & KIRC, LIHC, LSCC & LUAD, PAAD, PRAD, READ, THCA, LAML, STAD and SKCM). Thus, for each normal-adjacent sample of the TCGA we can estimate the total number of stem-cell divisions (TNSC) by multiplying the intrinsic rate (IR) of the tissue with the chronological age of the sample. Somatic mutational clock signature-1 (MS1) and signature-5 (MS5) were derived from Alexandrov et al [17]. These loads represent number of mutations per Mbp. Of note, because these mutational loads and stemTOC's mitotic-age both correlate with chronological age, in specific analyses we divide these values by the chronological age of the sample to arrive at age-adjusted values.

Code Availability

An R-package "EpiMitClocks" with functions to estimate the mitotic age for each of the clocks in this work is freely available from *https://github.com/aet21/EpiMitClocks* The package comes with a tutorial vignette. The R-package EpiSCORE for estimating cell-type fractions in solid tissue-types is available from *https://github.com/aet21/EpiSCORE*

Data Availability

The main Illumina DNA methylation datasets used here are freely available from public repositories, including GEO (www.ncbi.nlm.nih.gov/geo) and ArrayExpress (www.ebi.ac.uk/arrayexpress) (see Methods for relevant references). Details: Endicott (182 baseline-profiling samples from 6 cell lines, GEO: GSE197512); Hannum (656 whole blood, GEO: GSE40279); MESA (214 purified CD4+ T-cells and 1202 Monocyte samples, GEO: GSE56046 & GSE56581); Tserel (98 CD8+ T-cells, GEO: GSE59065); BLUEPRINT (139 matched CD4+ T-cells, Monocytes and Neutrophils, EGA: EGAS00001001456); Paul (100 B cells, 98 T cells, and 104 monocytes, EGA: EGAS00001001598) [82]; Gastric tissue (191 normal and metaplasia, GEO: GSE103186); Colon tissue (47 normal and adenoma, ArrayExpress: E-MTAB-6450); Colon tissue2

(123 normal, adenoma, and cancer samples, GEO: GSE48684) [97]; Breast Erlangen (397 normal, GEO: GSE69914); Breast2 (121 normal, GEO: GSE101961); Breast Johnson (55 normal and ductal carcinoma in situ samples, GEO: GSE66313); Liver tissue (137 normal, premalignant and cholangiocarcinoma samples, GEO: GSE156299); Prostate tissue (36 normal, neoplasia, primary and metastatic samples, GEO: GSE116338); Oral tissue (41 normal, lichen planus and oral squamous cell carcinoma samples, GEO: GSE123781); Esophagus (50 normal, 81 Barrett's Esophagus and 24 adenocarcinomas, GEO:GSE104707); Liver-NAFLD (n=325, GEO: GSE180474); SCM2 (37 fetal tissue samples, GEO: GSE31848); Cord-Blood (15 samples, GEO: GSE72867); Slieker (34 fetal tissue samples, GEO: GSE56515); eGTEX (987 samples from 9 normal tissue-types, GEO: GSE213478). TCGA data was downloaded from https://gdc.cancer.gov . The DNAm dataset in buccal cells from the NSHD MRC1946 is available by submitting data requests to mrclha.swiftinfo@ucl.ac.uk; see full policy at http://www.nshd.mrc.ac.uk/data.aspx. Managed access is in place for this 69-year-old study to ensure that use of the data is within the bounds of consent given previously by participants and to safeguard any potential threat to anonymity since the participants are all born in the same week. The DNAm-atlas encompassing DNAm reference matrices for 13 tissue-types encompassing over 40 cell-types is freely available from the EpiSCORE R-package https://github.com/aet21/EpiSCORE

Declarations

Ethics approval and consent to participate Not applicable here since we have only analyzed existing publicly available data.

Consent for publication: Not applicable.

Availability of data and materials: The main Illumina DNA methylation 450k/EPIC datasets used here are freely available from public repositories, including GEO (www.ncbi.nlm.nih.gov/geo), ArrayExpress (www.ebi.ac.uk/arrayexpress). Details: Endicott (182 cell-line samples, GEO: GSE197512) [30]. Hannum (656 whole blood, GEO: GSE40279) [71]; MESA (214 purified CD4+ T-cells and 1202 Monocyte samples, GEO: GSE56046 & GSE56581) [79]; Tserel (98 CD8+ T-cells, GEO: GSE59065) [80]; BLUEPRINT (139 matched CD4+ T-cells, Monocytes and Neutrophils, EGA: EGAS00001001456) [77]; Paul (100 B cells, 98 T cells, and 104 monocytes, EGA: EGAS00001001598) [82]; Liu (335 whole blood, GEO: GSE42861) [101]; Pai (n=28 sorted neurons, GEO: GSE112179) [83], Guintivano (n=29 sorted neurons, GEO: GSE41826) [84], Gasparoni (n=16 sorted neurons, GEO: GSE66351) [85] and Kozlenkov (n=29 sorted neurons, GEO: GSE98203) [86]; Gastric tissue (191 normal and metaplasia, GEO: GSE103186) [89]; Colon tissue (47 normal and adenoma, ArrayExpress: E-MTAB-6450)

[93]; Colon tissue2 (123 normal, adenoma, and cancer samples, GEO: GSE48684) [97]; Breast Erlangen (397 normal, GEO: GSE69914) [42]; Breast2 (121 normal, GEO: GSE101961) [102] ; Breast Johnson (55 normal and ductal carcinoma in situ samples, GEO: GSE66313) [87] ; Liver tissue (137 normal, premalignant and cholangiocarcinoma samples, GEO: GSE156299) [94] ; Prostate tissue (36 normal, neoplasia, primary and metastatic samples, GEO: GSE116338)[95]; Oral tissue (41 normal, lichen planus and oral squamous cell carcinoma samples, GEO: GSE123781) [96]; Esophagus (50 normal, 81 Barrett's Esophagus and 24 adenocarcinomas, GEO:GSE104707) [91]; Liver-NAFLD (n=325, GEO: GSE180474) [57]; SCM2 (37 fetal tissue samples, GEO: GSE31848) [64]; Cord-Blood (15 samples, GEO: GSE72867) [65]; Slieker (34 fetal tissue samples, GEO: GSE56515) [66]; eGTEX (987 samples from 9 normal tissue-types, GEO: GSE213478) [56]; TCGA data was downloaded from https://gdc.cancer.gov ; The DNAm dataset in buccal cells from the NSHD MRC1946 [54] is available by submitting data requests to mrclha.swiftinfo@ucl.ac.uk; see full policy at http://www.nshd.mrc.ac.uk/data.aspx. Managed access is in place for this 69-year-old study to ensure that use of the data is within the bounds of consent given previously by participants and to safeguard any potential threat to anonymity since the participants are all born in the same week. The DNAm-atlas encompassing DNAm reference matrices for 13 tissue-types encompassing over 40 cell-types is freely available from the EpiSCORE R-package https://github.com/aet21/EpiSCORE

Competing interests: The authors declare no competing interests.

Funding: This work was supported by NSFC (National Science Foundation of China) grants, grant numbers 32170652 and 31970632.

Authors contributions: AET conceived and designed the study. AET wrote the manuscript. TZ and AET performed the statistical and bioinformatic analyses. HT and ZD helped with bioinformatic analyses.

Acknowledgements: We would like to thank the TCGA and everyone who supports open-access data.

References

1. Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81 (2015).
2. Zhu, L. *et al.* Multi-organ Mapping of Cancer Risk. *Cell* **166**, 1132-1146 e7 (2016).
3. Klutstein, M., Moss, J., Kaplan, T. & Cedar, H. Contribution of epigenetic mechanisms to variation in cancer risk among tissues. *Proc Natl Acad Sci U S A* **114**, 2230-2234 (2017).
4. Wu, S., Powers, S., Zhu, W. & Hannun, Y.A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43-7 (2016).
5. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330-1334 (2017).

6. Tomasetti, C. & Vogelstein, B. Cancer risk: role of environment-response. *Science* **347**, 729-31 (2015).
7. Issa, J.P. Aging and epigenetic drift: a vicious cycle. *J Clin Invest* **124**, 24-9 (2014).
8. Ushijima, T. & Hattori, N. Molecular pathways: involvement of Helicobacter pylori-triggered inflammation in the formation of an epigenetic field defect, and its usefulness as cancer risk and exposure markers. *Clin Cancer Res* **18**, 923-9 (2012).
9. Ushijima, T. Epigenetic field for cancerization. *J Biochem Mol Biol* **40**, 142-50 (2007).
10. Issa, J.P. Epigenetic variation and cellular Darwinism. *Nat Genet* **43**, 724-6 (2011).
11. Herceg, Z. *et al.* Roadmap for investigating epigenome deregulation and environmental origins of cancer. *Int J Cancer* **142**, 874-882 (2018).
12. Lappalainen, T. & Greally, J.M. Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet* **18**, 441-451 (2017).
13. Baylin, S.B. & Ohm, J.E. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer* **6**, 107-16 (2006).
14. Feinberg, A.P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet* **7**, 21-33 (2006).
15. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci U S A* **110**, 1999-2004 (2013).
16. Alexandrov, L.B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622 (2016).
17. Alexandrov, L.B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-7 (2015).
18. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J.H.J. The central role of DNA damage in the ageing process. *Nature* **592**, 695-703 (2021).
19. Blackburn, E.H., Greider, C.W. & Szostak, J.W. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat Med* **12**, 1133-8 (2006).
20. Siegmund, K.D., Marjoram, P., Woo, Y.J., Tavare, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc Natl Acad Sci U S A* **106**, 4828-33 (2009).
21. Kim, J.Y., Tavare, S. & Shibata, D. Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions. *Proc Natl Acad Sci U S A* **102**, 17739-44 (2005).
22. Yang, Z. *et al.* Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol* **17**, 205 (2016).
23. Youn, A. & Wang, S. The MiAge Calculator: a DNA methylation-based mitotic age calculator of human tissue types. *Epigenetics* **13**, 192-206 (2018).
24. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet* **50**, 591-602 (2018).
25. Teschendorff, A.E. *et al.* Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med* **4**, 24 (2012).
26. Shen, S.Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579-583 (2018).
27. Wang, T. *et al.* Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. *Genome Biol* **18**, 57 (2017).

28. Field, A.E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Mol Cell* **71**, 882-895 (2018).
29. Yatabe, Y., Tavare, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc Natl Acad Sci U S A* **98**, 10839-44 (2001).
30. Endicott, J.L., Nolte, P.A., Shen, H. & Laird, P.W. Cell division drives DNA methylation loss in late-replicating domains in primary human cells. *Nat Commun* **13**, 6659 (2022).
31. Curtius, K. *et al.* A Molecular Clock Infers Heterogeneous Tissue Age Among Patients with Barrett's Esophagus. *PLoS Comput Biol* **12**, e1004919 (2016).
32. Luebeck, G.E. *et al.* Implications of Epigenetic Drift in Colorectal Neoplasia. *Cancer Res* **79**, 495-504 (2019).
33. Teschendorff, A.E., Zhu, T., Breeze, C.E. & Beck, S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol* **21**, 221 (2020).
34. Zhu, T. *et al.* A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. *Nat Methods* **19**, 296-306 (2022).
35. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
36. Nejman, D. *et al.* Molecular rules governing de novo methylation in cancer. *Cancer Res* **74**, 1475-83 (2014).
37. Salas, L.A. *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun* **13**, 761 (2022).
38. Teschendorff, A.E. & Relton, C.L. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* **19**, 129-147 (2018).
39. Breeze, C.E. *et al.* eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Rep* **17**, 2137-2150 (2016).
40. Breeze, C.E. *et al.* eFORGE v2.0: updated analysis of cell type-specific signal in epigenomic data. *Bioinformatics* **35**, 4767-4769 (2019).
41. Chen, Y., Breeze, C.E., Zhen, S., Beck, S. & Teschendorff, A.E. Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. *Epigenetics Chromatin* **9**, 10 (2016).
42. Teschendorff, A.E. *et al.* DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun* **7**, 10478 (2016).
43. Teschendorff, A.E. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome Med* **12**, 56 (2020).
44. Duran-Ferrer, M. *et al.* The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nat Cancer* **1**, 1066-1081 (2020).
45. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355-364 (2023).
46. Alcantara Llaguno, S.R. & Parada, L.F. Cell of origin of glioma: biological and clinical implications. *Br J Cancer* **115**, 1445-1450 (2016).
47. Xin, L. Cells of origin for cancer: an updated view from prostate cancer. *Oncogene* **32**, 3655-63 (2013).
48. Wang, X. *et al.* A luminal epithelial stem cell that is a cell of origin for prostate cancer. *Nature* **461**, 495-500 (2009).
49. Goldstein, A.S., Huang, J., Guo, C., Garraway, I.P. & Witte, O.N. Identification of a cell of origin for human prostate cancer. *Science* **329**, 568-71 (2010).

50. Zhang, D., Zhao, S., Li, X., Kirk, J.S. & Tang, D.G. Prostate Luminal Progenitor Cells in Development and Cancer. *Trends Cancer* **4**, 769-783 (2018).
51. Wang, Z.A., Toivanen, R., Bergren, S.K., Chambon, P. & Shen, M.M. Luminal cells are favored as the cell of origin for prostate cancer. *Cell Rep* **8**, 1339-46 (2014).
52. Karthaus, W.R. *et al.* Regenerative potential of prostate luminal cells revealed by single-cell analysis. *Science* **368**, 497-505 (2020).
53. Wattacheril, J.J., Raj, S., Knowles, D.A. & Greally, J.M. Using epigenomics to understand cellular responses to environmental influences in diseases. *PLoS Genet* **19**, e1010567 (2023).
54. Teschendorff, A.E. *et al.* Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA Oncol* **1**, 476-85 (2015).
55. Zheng, S.C. *et al.* A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics* **10**, 925-940 (2018).
56. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet* **55**, 112-122 (2023).
57. Johnson, N.D. *et al.* Differential DNA methylation and changing cell-type proportions as fibrotic stage progresses in NAFLD. *Clin Epigenetics* **13**, 152 (2021).
58. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-59 (2013).
59. Aran, D., Sirota, M. & Butte, A.J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971 (2015).
60. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640-646 (2020).
61. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).
62. Gabbutt, C. *et al.* Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nat Biotechnol* **40**, 720-730 (2022).
63. Teschendorff, A.E., Jones, A. & Widschwendter, M. Stochastic epigenetic outliers can define field defects in cancer. *BMC Bioinformatics* **17**, 178 (2016).
64. Nazor, K.L. *et al.* Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell* **10**, 620-34 (2012).
65. Aranyi, T. *et al.* Systemic epigenetic response to recombinant lentiviral vectors independent of proviral integration. *Epigenetics Chromatin* **9**, 29 (2016).
66. Slieker, R.C. *et al.* DNA Methylation Landscapes of Human Fetal Development. *PLoS Genet* **11**, e1005583 (2015).
67. Aryee, M.J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-9 (2014).
68. Teschendorff, A.E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-96 (2013).
69. You, C. *et al.* A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. *Nat Commun* **11**, 4779 (2020).
70. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
71. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* **49**, 359-67 (2013).

72. Johansson, A., Enroth, S. & Gyllensten, U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS One* **8**, e67378 (2013).
73. Zhou, W., Triche, T.J., Jr., Laird, P.W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res* **46**, e123 (2018).
74. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).
75. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416 e11 (2018).
76. Tian, Y. *et al.* ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982-3984 (2017).
77. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).
78. Teschendorff, A.E., Jing, H., Paul, D.S., Virta, J. & Nordhausen, K. Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biol* **19**, 76 (2018).
79. Reynolds, L.M. *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat Commun* **5**, 5366 (2014).
80. Tserel, L. *et al.* Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. *Sci Rep* **5**, 13107 (2015).
81. Zheng, S.C. *et al.* Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nat Methods* **14**, 216-217 (2017).
82. Paul, D.S. *et al.* Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat Commun* **7**, 13555 (2016).
83. Pai, S. *et al.* Differential methylation of enhancer at IGF2 is associated with abnormal dopamine synthesis in major psychosis. *Nat Commun* **10**, 2046 (2019).
84. Guintivano, J., Aryee, M.J. & Kaminsky, Z.A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290-302 (2013).
85. Gasparoni, G. *et al.* DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. *Epigenetics Chromatin* **11**, 41 (2018).
86. Kozlenkov, A. *et al.* DNA Methylation Profiling of Human Prefrontal Cortex Neurons in Heroin Users Shows Significant Difference between Genomic Contexts of Hyper- and Hypomethylation and a Younger Epigenetic Age. *Genes (Basel)* **8**(2017).
87. Johnson, K.C. *et al.* DNA methylation in ductal carcinoma in situ related with future development of invasive breast cancer. *Clin Epigenetics* **7**, 75 (2015).
88. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-5 (2001).
89. Huang, K.K. *et al.* Genomic and Epigenomic Profiling of High-Risk Intestinal Metaplasia Reveals Molecular Determinants of Progression to Gastric Cancer. *Cancer Cell* **33**, 137-150 e5 (2018).
90. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
91. Luebeck, E.G. *et al.* Identification of a key role of widespread epigenetic drift in Barrett's esophagus and esophageal adenocarcinoma. *Clin Epigenetics* **9**, 113 (2017).
92. Maity, A.K. *et al.* Novel epigenetic network biomarkers for early detection of esophageal cancer. *Clin Epigenetics* **14**, 23 (2022).
93. Bormann, F. *et al.* Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal

Carcinogenesis. *Cell Rep* **23**, 3407-3418 (2018).
94. Goeppert, B. *et al.* Integrative analysis reveals early and distinct genetic and epigenetic changes in intraductal papillary and tubulopapillary cholangiocarcinogenesis. *Gut* **71**, 391-401 (2022).
95. Silva, R. *et al.* Longitudinal analysis of individual cfDNA methylome patterns in metastatic prostate cancer. *Clin Epigenetics* **13**, 168 (2021).
96. Nemeth, C.G. *et al.* Recurrent chromosomal and epigenetic alterations in oral squamous cell carcinoma and its putative premalignant condition oral lichen planus. *PLoS One* **14**, e0215055 (2019).
97. Luo, Y. *et al.* Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* **147**, 418-29 e8 (2014).
98. Luo, Q. *et al.* A meta-analysis of immune cell fractions at high resolution reveals novel associations with common phenotypes and health outcomes. *biorxiv* **https://doi.org/10.1101/2023.03.20.533349**(2023).
99. Teschendorff, A.E., Breeze, C.E., Zheng, S.C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).
100. Teschendorff, A.E. & Widschwendter, M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487-94 (2012).
101. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **31**, 142-7 (2013).
102. Song, M.A. *et al.* Landscape of genome-wide age-related DNA methylation in breast tissue. *Oncotarget* **8**, 114648-114662 (2017).

Figure Legends

Figure-1: Construction of stemTOC and estimation of mitotic age. a) We first identify CpGs (n=30,257) mapping to within 200bp upstream of the TSS of genes that are unmethylated (defined by DNAm beta-value < 0.2) across 86 fetal-tissue samples from 13 different fetal-tissues. These are then filtered further by the requirements that they display hypermethylation as a function of population doublings in cell-lines, and that they don't display such hypermethylation when cell-lines are deprived from growth-promoting serum or when treated with mitomycin (MMC, a cell-cycle inhibitor), resulting in 629 CpGs satisfy these conditions. **b)** To exclude cell-culture specific effects, we then ask if DNAm at each of these 629 CpGs displays significant hypermethylation with chronological age, as assessed in 3 separate whole blood cohorts, after adjusting for variations in 12 immune-cell type fractions. This results in 371 "vivo-mitCpGs". **c)** Simulation model of DNAm changes at 20 CpGs during carcinogenic transformation (a total of 40 independent samples, 10 from each disease-stage as shown) . In the initial stages, DNAm changes are inherently stochastic, which means

that an average DNAm over the CpGs may not discriminate normal-healthy from normal "at-risk" tissue. Instead, taking an upper-quantile (in this case the upper 95% quantile) of the DNAm value over the CpGs can discriminate normal from normal at-risk tissue. **d)** Top: Scatterplot depicts the difference in upper quantiles (y-axis) over the 371 vivo-mitCpGs between 42 normal samples adjacent to breast cancer (NADJ, "normal at-risk") and 50 age-matched healthy breast samples (N, "normal healthy"), against the upper quantile threshold (x-axis). Horizontal blue line indicates the difference in mean DNAm over the 371 CpGs between NADJ and N tissue. Vertical red line marks the selected upper-quantile threshold maximizing difference without compromising variability. A relative mitotic-age score (stemTOC) is obtained for any sample, by taking the 95% upper quantile (UQ) over the 371 DNAm beta-values corresponding to these vivo-mitCpGs. Bottom density plots display the distribution of the 371 DNAm values for one hypothetical normal healthy and one hypothetical normal at-risk sample. Vertical lines indicate the stemTOC-scores, i.e. the 95% upper quantiles defined over the 371 CpGs.

Figure-2: Validation and benchmarking of stemTOC. a) For selected normal-adjacent tissue-types, scatterplots of stemTOC's mitotic age (y-axis) vs chronological age (x-axis). Number of normal-adjacent samples is given above each plot. The Pearson Correlation Coefficient (PCC) and P-value from a linear regression is given. Normal-adjacent tissue-types in TCGA are labelled by the corresponding cancer-type. **b)** Heatmap of PCC values between mitotic and chronological age for 7 mitotic clocks and across all normal-adjacent tissue-types from the TCGA. **c)** As a), but for three sorted immune-cell populations as profiled by BLUEPRINT. **d)** As b) but for all sorted immune-cell populations. Sorted immune-cell samples labeled by cell-type and study it derives from. BP=BLUEPRINT. **e)** As a) but for sorted neurons from 4 different cohorts, indicated by color. In this panel the PCC is a weighted average over the 4 cohorts and the P-value is adjusted for cohort. **f)** Scatterplot of a normal tissue's relative intrinsic rate (RIR) estimated from taking the median of stemTOC's mitotic age divided by chronological age and multiplied by 10 (y-axis), against the Vogelstein-Tomasetti intrinsic annual rate of stem-cell division (IR) (x-axis), using the normal-adjacent tissues with such estimates. The Pearson and Spearman Correlation Coefficients (PCC & SCC) are given. **g)** Top: Scatterplot of a normal sample's mitotic age (estimated with stemTOC) vs the estimated total number of stem-cell divisions (TNSC=IR*Age) expressed in a log2-basis, to better highlight the differences between the low and medium turnover tissues. P-value is from a multivariate linear regression including age as a covariate. Bottom: Violin plot representation of middle panel, with normal-adjacent tissue samples grouped into 5 IR-classes, as shown. Low: Thyroid+Lung, MedLow=Liver+Pancreas+Kidney, Med=Prostate+Breast+Bladder, MedHigh=Esophagus+Oral, High=Colon+Rectum+Stomach. **h)** Barplots compare the PCC and significance level attained by stemTOC shown in panel e) to the corresponding values for 6 other mitotic clocks. Mitotic clocks based on CpGs that gain/lose methylation with cell-division (hypermethylated/hypomethylated) are

indicated.

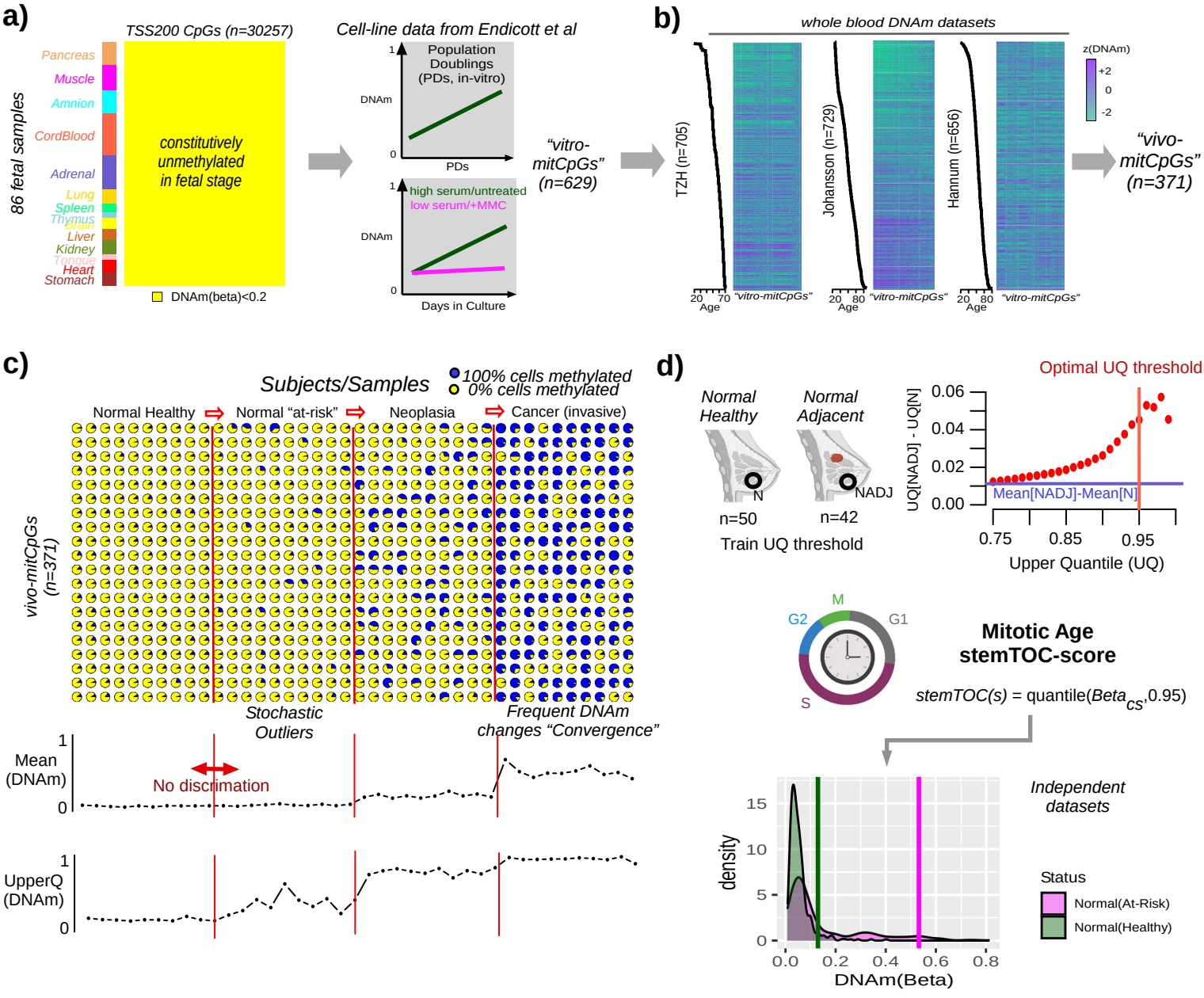
Figure-3: Correlation of stemTOC with tumor cell-of-origin fraction. a) Alluvial plot displaying the validation of EpiSCORE on sorted WGBS samples from Loyfer et a DNAm-atlas. The left and middle bar label the tissue-type and cell-type of the WGBS sample from Loyfer et al DNAm-atlas. Right bar labels the predicted cell-type using EpiSCORE. Overall prediction accuracy is given below. **b)** Scatterplot of stemTOC's relative mitotic age estimate (y-axis) vs. the presumed cell-of-origin fraction derived with EpiSCORE (x-axis) for selected TCGA cancer-types. The number of tumor samples is indicated above the plot. For each panel we provide the Pearson Correlation Coefficient (PCC) and P-value from a linear regression. Of note, in each panel we are displaying the cell-type that displayed the strongest correlation with mitotic-age and this coincides with the presumed cell-of-origin. **c)** As b), but for 3 tumor-types where the tumor cell-of-origin is less well established or controversial. **d)** Heatmap of Pearson correlations for all 7 mitotic clocks and across all 16 cancer-types. **e)** Heatmap displaying one-tailed paired Wilcoxon rank sum test P-values, comparing clocks to each other, in how well their mitotic age correlates with the tumor cell-of-origin fraction. Each row indicates how well the corresponding clock's mitotic age estimate performs in relation to the clock specified by the column. The paired Wilcoxon test is performed over the 16 cancer-types.

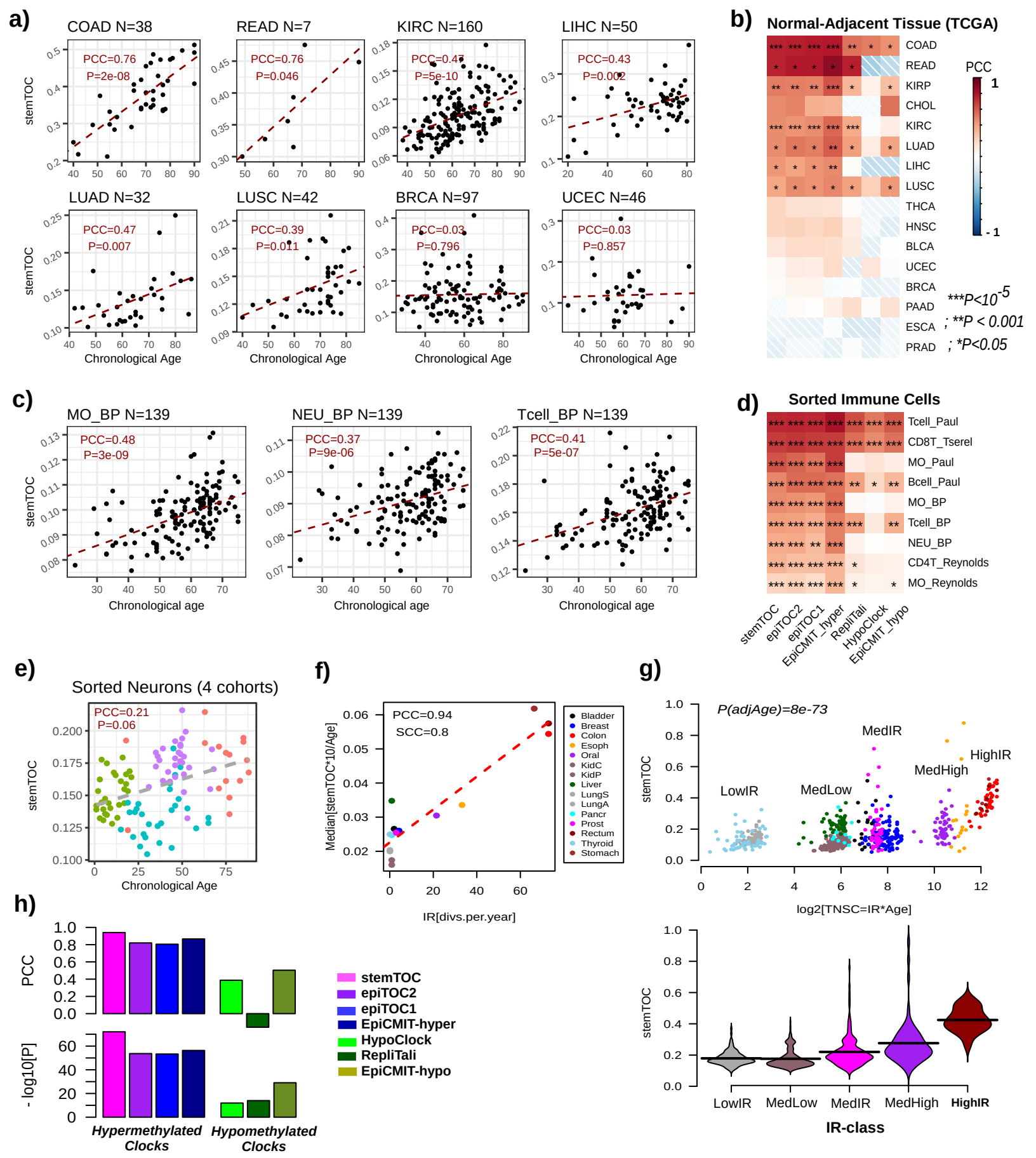
Figure-4: stemTOC correlates with cancer risk and cancer-risk factors. a) Violin plots displaying the distribution of stemTOC's mitotic age for normal tissue (N) and age-matched precancerous (PreC) samples for each of 9 different DNAm datasets. Number of samples in each category is displayed below x-axis, as well as the tissue-type and the nature of the precancerous sample. DCIS=ductal carcinoma in-situ, PreM=premalignant, OLP=oral lichen planus, LCIS=lung carcinoma in-situ. P-value derives from a one-tailed Wilcoxon rank sum test. **b)** Top: Violin plots depicting the estimated epithelial and immune cell fraction in 790 buccal swab samples from women all aged-53 at sample draw. Bottom left: Scatterplot of stemTOC's mitotic age (y-axis) against estimated squamous epithelial fraction (x-axis) for the same 790 buccal swab samples. Samples are colored by their lifelong smoking habit (current, ex-smoker and never-smoker) with color indicated in panel to the right. P-value is from a multivariate linear regression including epithelial fraction and smoking-status. Bottom right: Violin plots displaying the distribution of stemTOC's mitotic age as a function of smoking-status, as indicated. P-value is from a multivariate linear regression including smoking-status and squamous epithelial fraction. **c)** Top: Violin plots displaying the estimated cell-type fraction for 7 lung cell-types (Endo=endothelial, Epi=epithelial, Gran=granulocyte, Lym=lymphocyte, Macro=macrophage, Mono=monocyte, Stromal) in 204 normal lung samples from eGTEX. Bottom: Violin plots displaying stemTOC's mitotic age as a function of chronological age and smoking status in the same 204 normal lung-tissue samples from eGTEX. Number of samples in each smoking category and age-group is shown below each box. Violin colors label smoking-status

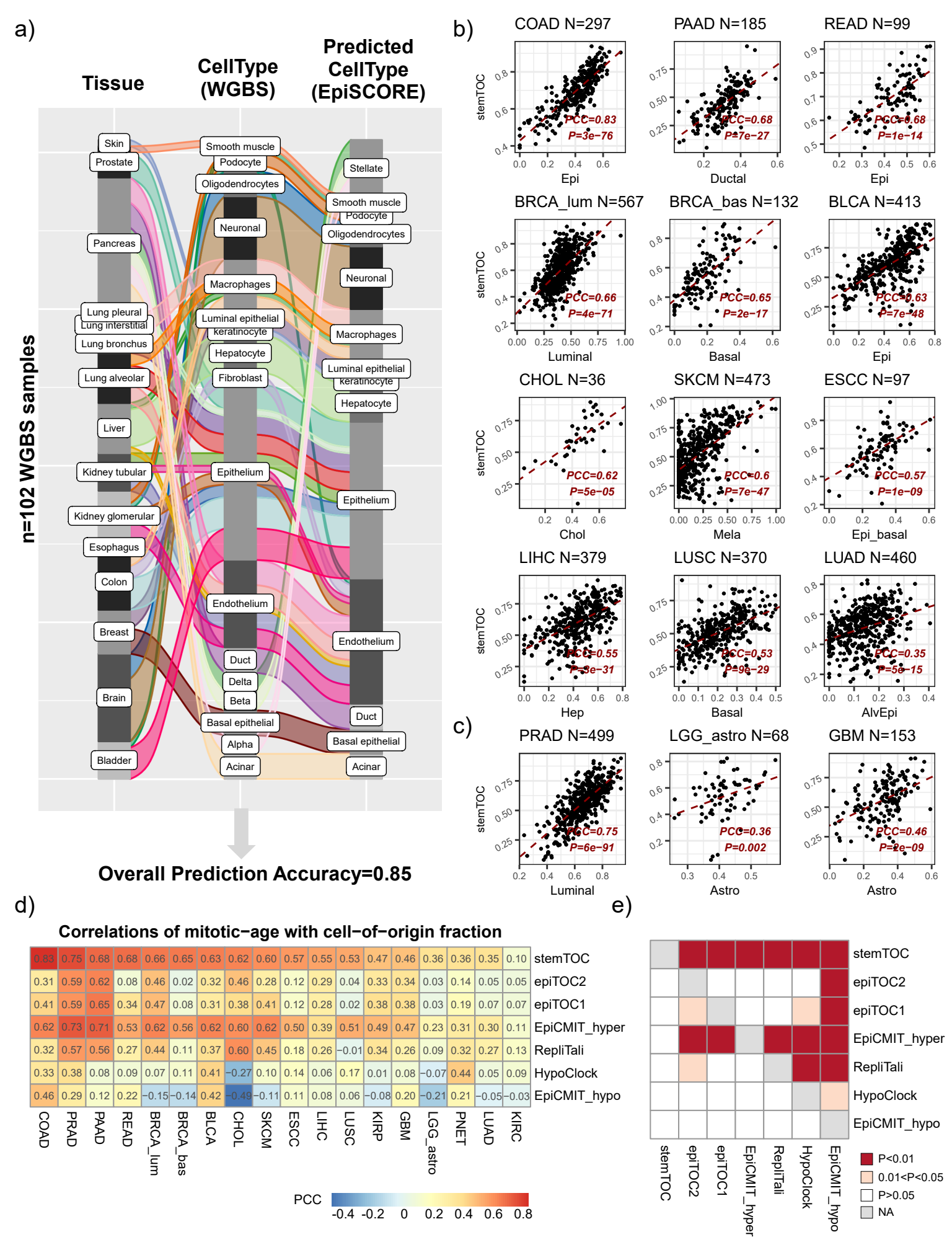
as in b). P-values for age and smoking status are derived from a multivariate linear regression that included age, smoking and lung epithelial fraction. **d)** Top left: Violin plots of estimated cell-type fractions for 5 liver cell-types (Hep=hepatocytes, Chol=cholangiocytes, EC=endothelial cells, Kup=Kupffer cells, Lym=lymphocytes) in 325 liver samples. Top right: Violin plots of the hepatocyte fraction against NAFLD disease stage. Number of samples in each group is shown and P-value is from a linear regression. Middle: Scatterplot of stemTOC's mitotic age (y-axis) against chronological age (x-axis) for the same 325 liver samples, with colors labeling the stage of NAFLD, with color label as in previous panel. Bottom: Violin plots display the distribution of mitotic age against NAFLD stage. P-value for disease stage derives from a multivariate linear regression that included age and hepatocyte fraction besides NAFLD-stage.

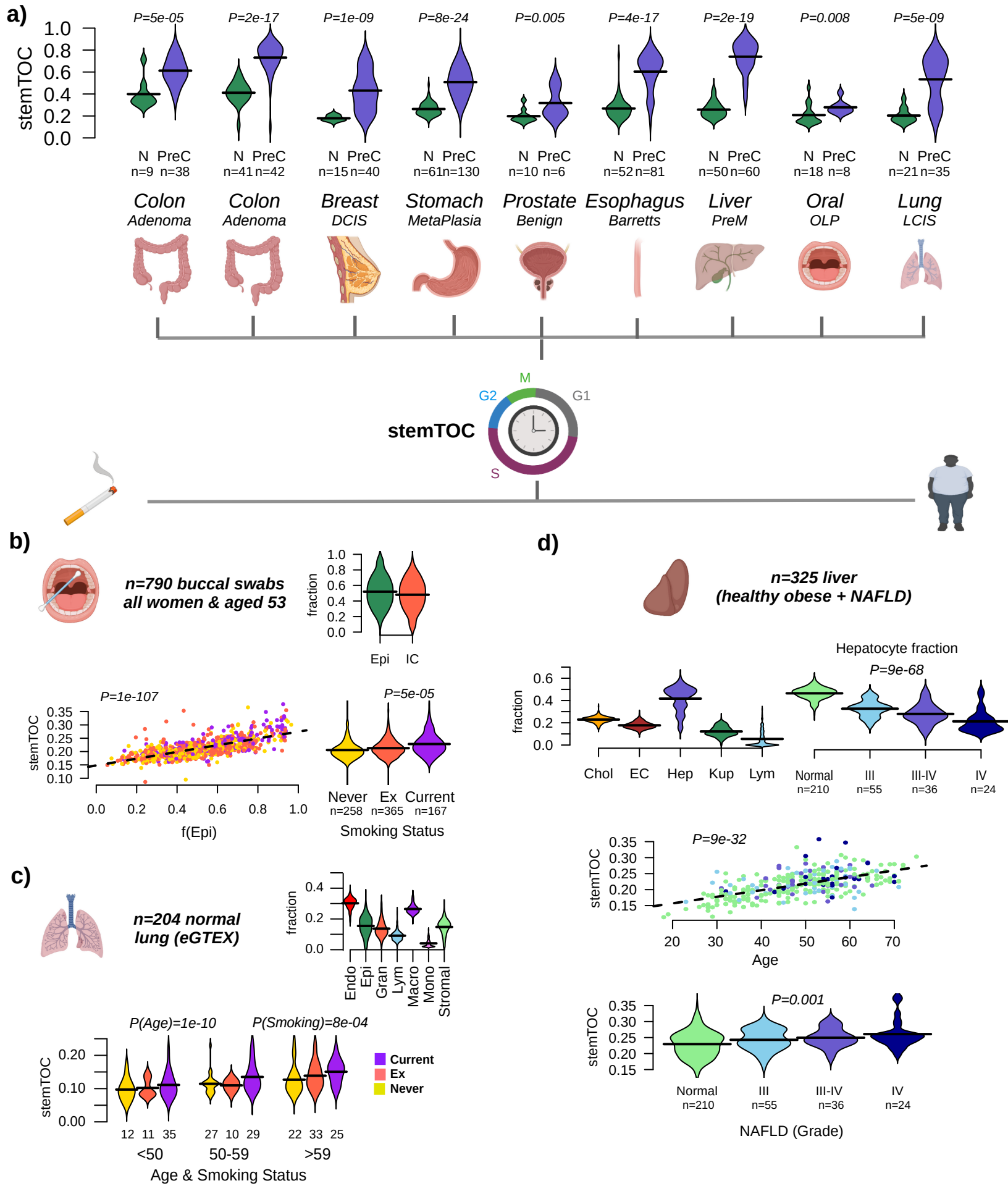
Figure-5: Relation of stemTOC with somatic mutational MS1 and MS5 signatures. **a)** Scatterplot of MS1 signature load (Mutations/Mb, y-axis) against the estimated total number of stem-cell divisions (Age*intrinsic annual rate of stem-cell division of the corresponding normal-tissue (IR), x-axis) for >7000 cancer samples from 17 TCGA cancer-types. Only cancer-types with an IR estimate in normal tissue were used. P-values derive from a multivariate linear regression including age as a covariate. **b)** As a) but for somatic mutational signature MS5. **c)** Median value of age-adjusted MS1-signature load (x10) for each cancer-type vs the annual intrinsic rate of stem-cell division of the corresponding normal tissue-type. Both Pearson (PCC) and Spearman (SCC) correlation coefficients are given. P-value tests for significance of SCC. **d)** As c) but for MS5. **e)** As a), but for stemTOC's mitotic age. **f)** As a), but for stemTOC, adjusted for chronological age and multiplied by 10 to reflect DNAm change over a decade. **g)** Scatterplot of the median stemTOC value of each TCGA cancer-type (y-axis) vs the median mutational load (mutations per Mb) of that cancer-type, for MS1. Spearman's correlation coefficient and P-value are given. **h)** As g) but for MS5.

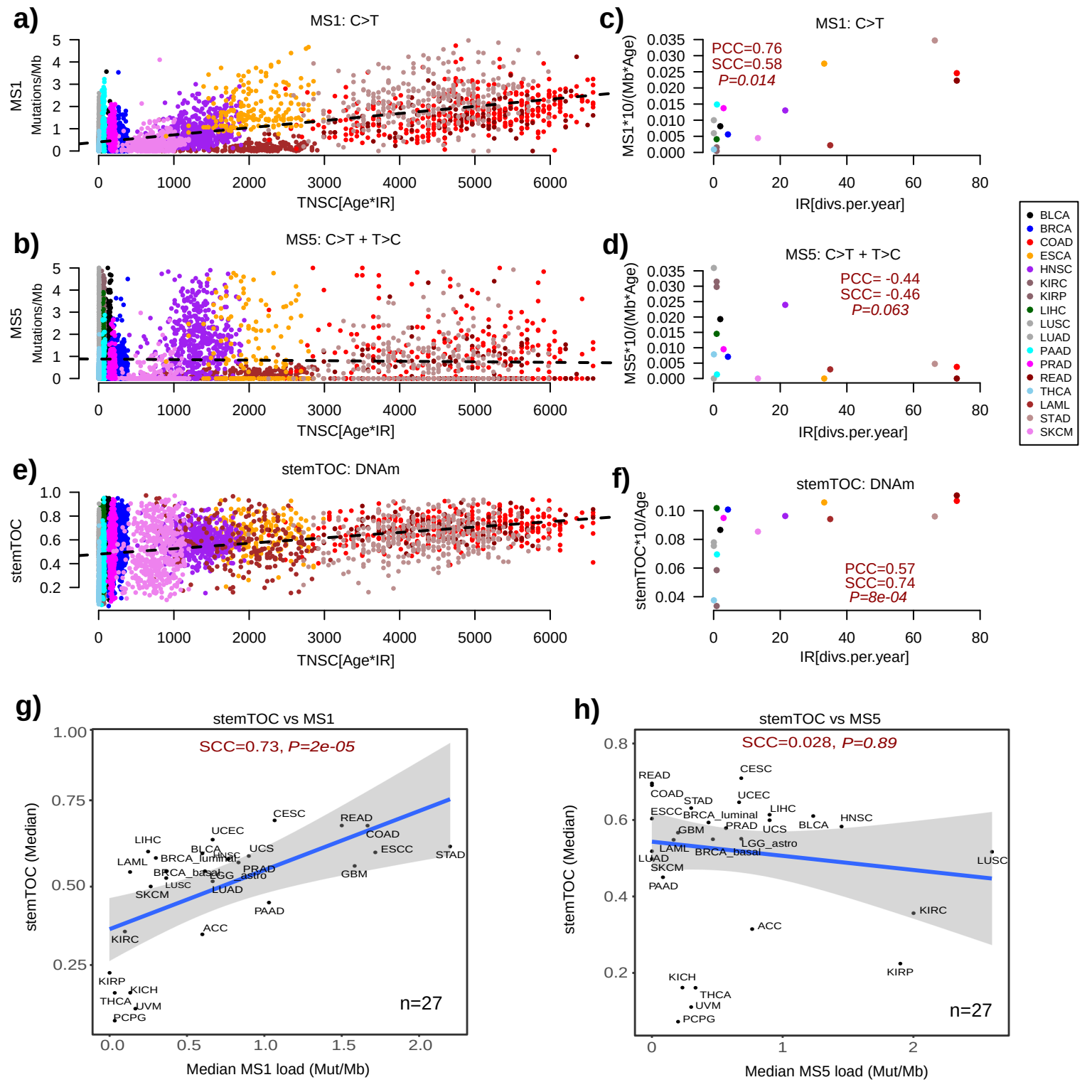
Construction of the stochastic epigenetic mitotic timer of cancer (stemTOC)




a)
COAD N=38
PCC=0.76
P=2e-08
READ N=7
PCC=0.76
P=0.046
KIRC N=160
PCC=0.43
P=5e-10
LIHC N=50
PCC=0.43
P=0.002
LUAD N=32
PCC=0.47
P=0.007
LUSC N=42
PCC=0.39
P=0.011
BRCA N=97
PCC=0.03
P=0.796
UCEC N=46
PCC=0.03
P=0.857
stemTOC
Chronological Age
b)
Normal-Adjacent Tissue (TCGA)
COAD
READ
KIRP
CHOL
KIRC
LUAD
LIHC
LUSC
THCA
HNSC
BLCA
UCEC
BRCA
PAAD
ESCA
PRAD
PCC
1
- 1
***P<10-5 ; **P < 0.001 ; *P<0.05
c)
MO_BP N=139
PCC=0.48
P=3e-09
NEU_BP N=139
PCC=0.37
P=9e-06
Tcell_BP N=139
PCC=0.41
P=5e-07
stemTOC
Chronological age
d)
Sorted Immune Cells
Tcell_Paul
CD8T_Tserel
MO_Paul
Bcell_Paul
MO_BP
Tcell_BP
NEU_BP
CD4T_Reynolds
MO_Reynolds
stemTOC
epiTOC2
epiTOC1
EpiCMIT_hyper
RepliTali
HypoClock
EpiCMIT_hypo
e)
Sorted Neurons (4 cohorts)
PCC=0.21
P=0.06
stemTOC
Chronological Age
f)
PCC=0.94
SCC=0.8
Median[stemTOC*10/Age]
IR[divs.per.year]
Bladder
Breast
Colon
Esoph
Oral
KidC
KidP
Liver
LungS
LungA
Pancr
Prost
Rectum
Thyroid
Stomach
g)
P(adjAge)=8e-73
LowIR
MedLow
MedIR
MedHigh
HighIR
stemTOC
log2[TNSC=IR*Age]
IR-class
h)
PCC
- log10[P]
stemTOC
epiTOC2
epiTOC1
EpiCMIT-hyper
HypoClock
RepliTali
EpiCMIT-hypo
Hypermethylated Clocks
Hypomethylated Clocks


a)
Tissue
CellType (WGBS)
Predicted CellType (EpiSCORE)
n=102 WGBS samples
Skin
Prostate
Pancreas
Lung pleural
Lung interstitial
Lung bronchus
Lung alveolar
Liver
Kidney tubular
Kidney glomerular
Esophagus
Colon
Breast
Brain
Bladder
Smooth muscle
Podocyte
Oligodendrocytes
Neuronal
Macrophages
Luminal epithelial
keratinocyte
Hepatocyte
Fibroblast
Epithelium
Endothelium
Duct
Delta
Beta
Basal epithelial
Alpha
Acinar
Stellate
Smooth muscle
Podocyte
Oligodendrocytes
Neuronal
Macrophages
Luminal epithelial
keratinocyte
Hepatocyte
Epithelium
Endothelium
Duct
Basal epithelial
Acinar
Overall Prediction Accuracy=0.85
b)
COAD N=297
PAAD N=185
READ N=99
BRCA_lum N=567
BRCA_bas N=132
BLCA N=413
CHOL N=36
SKCM N=473
ESCC N=97
LIHC N=379
LUSC N=370
LUAD N=460
stemTOC
Epi
Ductal
Luminal
Basal
Chol
Mela
Epi_basal
Hep
AlvEpi
PCC=0.83
P=3e−76
PCC=0.68
P=7e−27
PCC=0.68
P=1e−14
PCC=0.66
P=4e−71
PCC=0.65
P=2e−17
PCC=0.63
P=7e−48
PCC=0.62
P=5e−05
PCC=0.6
P=7e−47
PCC=0.57
P=1e−09
PCC=0.55
P=3e−31
PCC=0.53
P=9e−29
PCC=0.35
P=5e−15
c)
PRAD N=499
LGG_astro N=68
GBM N=153
Luminal
Astro
PCC=0.75
P=6e−91
PCC=0.36
P=0.002
PCC=0.46
P=2e−09
d)
Correlations of mitotic−age with cell−of−origin fraction
stemTOC 0.83 0.75 0.68 0.68 0.66 0.65 0.63 0.62 0.60 0.57 0.55 0.53 0.47 0.46 0.36 0.36 0.35 0.10
epiTOC2 0.31 0.59 0.62 0.08 0.46 0.02 0.32 0.46 0.28 0.12 0.29 0.04 0.33 0.34 0.03 0.14 0.05 0.05
epiTOC1 0.41 0.59 0.65 0.34 0.47 0.08 0.31 0.38 0.41 0.12 0.28 0.02 0.38 0.38 0.03 0.19 0.07 0.07
EpiCMIT_hyper 0.62 0.73 0.71 0.53 0.62 0.56 0.62 0.60 0.62 0.50 0.39 0.51 0.49 0.47 0.23 0.31 0.30 0.11
RepliTali 0.32 0.57 0.56 0.27 0.44 0.11 0.37 0.60 0.45 0.18 0.26 −0.01 0.34 0.26 0.09 0.32 0.27 0.13
HypoClock 0.33 0.38 0.08 0.09 0.07 0.11 0.41 −0.27 0.10 0.14 0.06 0.17 0.01 0.08 −0.07 0.44 0.05 0.09
EpiCMIT_hypo 0.46 0.29 0.12 0.22 −0.15 −0.14 0.42 −0.49 −0.11 0.11 0.08 0.06 −0.08 0.20 −0.21 0.21 −0.05 −0.03
COAD PRAD PAAD READ BRCA_lum BRCA_bas BLCA CHOL SKCM ESCC LIHC LUSC KIRP GBM LGG_astro PNET LUAD KIRC
PCC
-0.4 -0.2 0 0.2 0.4 0.6 0.8
e)
stemTOC
epiTOC2
epiTOC1
EpiCMIT_hyper
RepliTali
HypoClock
EpiCMIT_hypo
P<0.01
0.01<P<0.05
P>0.05
NA


a)
stemTOC
1.0
0.8
0.6
0.4
0.2
0.0
P=5e-05
P=2e-17
P=1e-09
P=8e-24
P=0.005
P=4e-17
P=2e-19
P=0.008
P=5e-09
N PreC
n=9 n=38
n=41 n=42
n=15 n=40
n=61 n=130
n=10 n=6
n=52 n=81
n=50 n=60
n=18 n=8
n=21 n=35
Colon Adenoma
Colon Adenoma
Breast DCIS
Stomach MetaPlasia
Prostate Benign
Esophagus Barretts
Liver PreM
Oral OLP
Lung LCIS
stemTOC
M
G2
G1
S
b)
n=790 buccal swabs
all women & aged 53
fraction
Epi
IC
P=1e-107
stemTOC
f(Epi)
P=5e-05
Never
Ex
Current
n=258
n=365
n=167
Smoking Status
c)
n=204 normal
lung (eGTEX)
fraction
Endo
Epi
Gran
Lym
Macro
Mono
Stromal
P(Age)=1e-10
P(Smoking)=8e-04
stemTOC
12 11 35
27 10 29
22 33 25
<50
50-59
>59
Age & Smoking Status
Current
Ex
Never
d)
n=325 liver
(healthy obese + NAFLD)
Hepatocyte fraction
P=9e-68
fraction
Chol
EC
Hep
Kup
Lym
Normal
n=210
III
n=55
III-IV
n=36
IV
n=24
P=9e-32
stemTOC
Age
P=0.001
stemTOC
Normal
n=210
III
n=55
III-IV
n=36
IV
n=24
NAFLD (Grade)


a)
MS1: C>T
MS1
Mutations/Mb
TNSC[Age*IR]
b)
MS5: C>T + T>C
MS5
Mutations/Mb
TNSC[Age*IR]
c)
MS1: C>T
PCC=0.76
SCC=0.58
P=0.014
MS1*10/(Mb*Age)
IR[divs.per.year]
d)
MS5: C>T + T>C
PCC= -0.44
SCC= -0.46
P=0.063
MS5*10/(Mb*Age)
IR[divs.per.year]
BLCA
BRCA
COAD
ESCA
HNSC
KIRC
KIRP
LIHC
LUSC
LUAD
PAAD
PRAD
READ
THCA
LAML
STAD
SKCM
e)
stemTOC: DNAm
stemTOC
TNSC[Age*IR]
f)
stemTOC: DNAm
PCC=0.57
SCC=0.74
P=8e-04
stemTOC*10/Age
IR[divs.per.year]
g)
stemTOC vs MS1
SCC=0.73, P=2e-05
stemTOC (Median)
Median MS1 load (Mut/Mb)
n=27
h)
stemTOC vs MS5
SCC=0.028, P=0.89
stemTOC (Median)
Median MS5 load (Mut/Mb)
n=27