Age Influences on the Molecular Presentation of Tumours

Constance H. Li[1,2,3,4,5,6], Syed Haider[7], Paul C. Boutros[2,3,4,5,6,8,9]

[1]Ontario Institute for Cancer Research, Toronto, Ontario, Canada
[2] Department of Medical Biophysics, University of Toronto; Toronto, Ontario, Canada
[3] Department of Human Genetics, University of California, Los Angeles, CA, USA
[4] Department of Urology, University of California, Los Angeles, CA, USA
[5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, USA
[6] Institute for Precision Health, University of California, Los Angeles, CA, USA
[7]The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research; London, United Kingdom
[8]Department of Pharmacology & Toxicology, University of Toronto; Toronto, Ontario, Canada
[9]Vector Institute for Artificial Intelligence, Toronto, Canada

Correspondence should be addressed to P.C.B. (PBoutros@mednet.ucla.edu)
12-109 CHS
10833 Le Conte Avenue
Los Angeles, CA, 90095
Phone: 310-794-7160

Abstract

Cancer is often called a disease of aging. There are numerous ways in which cancer epidemiology and behaviour change with the age of the patient. The molecular bases for these relationships remain largely underexplored. To characterize them, we analyzed age-associations in the nuclear and mitochondrial somatic mutational landscape of 20,033 tumours across 35 tumour-types. Age influences both the number of mutations in a tumour and their evolutionary timing. Specific mutational signatures are associated with age, reflecting differences in exogenous and endogenous oncogenic processes such as a greater influence of tobacco use in the tumours of younger patients, but more activity of DNA damage repair signatures in those of older patients. We find that known cancer driver genes are mutated in age-associated frequencies, and these alter the transcriptome and predict for clinical outcomes. These effects are most striking in brain cancers where alterations like *SUFU* loss and *ATRX* mutation are age-dependent prognostic biomarkers. Using three cancer datasets, we show that age shapes the somatic mutational landscape of cancer, with clinical implications.

Introduction

Cancer health disparities across different population stratifiers are common through a wide range of measures. These include differences in incidence rates, mortality rates, response to treatment, and survival between individuals of different sexes[1–6], races or ancestries[7–11] and ages[12–14], and these differences have been described across a range of tumour-types. Cancer disparities involving age are particularly well known. Aging is a leading risk factor for cancer, as it is associated with increased incidence of most tumour-types[9,15]. Older age is also typically associated with higher mortality and lower survival[16,17]. The links between older age and increased cancer burden are such that cancer is often described as a disease of aging[18,19].

However, there are many nuances in the relationship between aging and cancer. Pediatric cancers are an obvious exception, as cancers arising in children have different molecular and clinical characteristics[9,20–22]. Tumours arising in young adults (< 50 years of age) are often more aggressive: early onset tumours of the prostate[23], breast[24], pancreatic[25,26], colorectal[27], and soft tissue sarcomas[28] are diagnosed at higher stages and associated with lower survival. Molecular studies have described some striking differences in the mutational landscapes of early onset *vs*. later onset disease[29–31], suggesting differences in the underlying oncogenic processes driving cancer at different ages.

The mechanisms of how age shapes the clinical behaviour of cancers have been subject to intense study. Many factors and behaviours closely tied to aging have been implicated in observed epidemiological and clinical cancer health disparities. For example, higher age is associated with a greater burden of comorbidities such as diabetes and cardiovascular disease[32,33]. Higher prevalence of chronic disease, frailty and increased likelihood of adverse drug reactions also influence the choices of clinical interventions given to older cancer patients[34–36]. Nevertheless differences remain even after accounting for these factors[37]. Previous work associating somatic molecular changes with age suggest differences in overall tumour mutation burden[38], transcriptional profiles[39], and some mutational differences[29–31]. These studies have focused on single tumour-types, relatively small cohorts, or have only evaluated fractions of the whole-genome, leaving the landscape of age-associated cancer mutations largely unknown.

To fill this gap, we performed a pan-cancer, genome-wide study of age-associated molecular differences in 10,218 tumours of 23 tumour-types from The Cancer Genome Atlas (TCGA), 2,562 tumours of 30 tumour-types from the International Cancer Genome Consortium/The Cancer Genome Atlas Pan-cancer Analysis of Whole Genomes (PCAWG), and 7,259 tumours across 35 tumour-types from the AACR Genomics Evidence Neoplasia Information Exchange (GENIE) projects. We quantified age-associations in measures of mutation density, subclonal architecture, mutation timing,

mutational signatures and driver mutations in almost all tumour-types. These associations remained even after adjusting for potential confounding factors such as sex and ancestry. Many of these genomic age-associations were linked to clinical phenotypes: in particular, we identified genomic alterations that were prognostic in specific age contexts, suggesting the clinical utility of age-informed biomarkers.

Results

Age Associations in Mutation Density and Timing

We investigated TCGA, PCAWG and AACR GENIE datasets independently and performed pan-cancer analyses spanning all TCGA, all PCAWG, and all AACR GENIE tumours in separate analyses; these were supplemented with tumour-type-specific analyses. We used the recorded age at diagnosis (**Table 1**). We used a two-stage statistical approach: we first used univariate methods to identify putative age-associations, and then further modeled these putative hits with multivariate regression to evaluate age effects after adjusting for confounding factors. Our multivariate modeling accounted for a range of confounding variables for each tumour-type including sex and genetic ancestry. We modeled each genomic feature and tumour subtype based on available clinical data, *a priori* knowledge, variable collinearity and model convergence. Model and variable specifications, and results of association tests between model variables and age are presented in **Supplementary Table 1**. We perform multiple testing adjustment at both stages using the Benjamini-Hochberg false discovery rate procedure and these adjusted p-values are used throughout. Bonferroni-adjusted p-values provide similar support for our findings. Both Benjamini-Hochberg and Bonferroni, as well as unadjusted p-values are provided in supplementary materials.

The accumulation of mutations with age is a well-known phenomenon in both cancer and non-cancer cells[41–48]. To test the robustness of our statistical framework in detecting age-associated genomic events, we investigated age associations in two measures of mutation accumulation: single nucleotide variant (SNV) density and genome instability. Both SNV density and genome instability have clinical relevance as they are associated with poor outcome in some tumour-types[49–51] and response to immunotherapy in others[52,53]. We first identified univariate age-associations in SNV density using Spearman correlation. Putative age-associations identified at a false discovery rate (FDR) threshold of 10% were further analysed by multivariate linear regression (LNR) models to adjust for tumour-type-specific confounding effects (**Supplementary Table 1**). We have previously applied this statistical strategy successfully to identify sex-associated somatic mutational features[40,54].

As expected[41,42], we identified significant associations between age and the accumulation of SNVs across a range of tumour contexts (**Figure 1A**). There were pan-cancer positive correlations between age and SNV density in TCGA (pan-TCGA: $\rho = 0.31$, adjusted LNR $p = 4.1 \times 10^{-57}$) and PCAWG ($\rho = 0.43$, adjusted LNR $p = 1.6 \times 10^{-26}$) data. Using TCGA and PCAWG data, we estimate that SNV density increases at a rate of 0.077 mutations per megabase pair per year (**Table 2**, **Methods**). We also identified positive associations in 11 TCGA, 14 PCAWG, and six AACR GENIE tumour-types (**Figure 1A**). Of these, eight tumour-types showed consistent results in two of three datasets (**Supplementary Figure**

1, Supplementary Table 2) including prostate cancer (TCGA: ρ = 0.25, adjusted LNR p = 0.015; PCAWG: ρ = 0.48, adjusted LNR p = 1.2×10^{-4}; **Figure 1B**). Estimates for per year increase in mutation density are given in **Table 2** for tumour-types with consistent evidence in at least two datasets.

We next asked whether there were differences in the timing of when these SNVs occurred during tumour evolution and leveraged data describing the evolutionary history of PCAWG tumours[55]. We first investigated polyclonality, or the number of cancer cell populations detected in each tumour. Monoclonal tumours, or those where all tumour cells are derived from one ancestral cell, are associated with better survival in several tumours types[56–58]. While there were intriguing univariate associations between age and polyclonality in non-Hodgkin lymphoma and prostate cancer, these were not significant after multivariate adjustment (**Supplementary Figure 1, Supplementary Table 2**). We then focused on polyclonal tumours and asked whether there were associations in mutation timing: we investigated whether SNVs, indels or structural variants (SVs) occurred more frequently as clonal mutations in the trunk or as subclonal ones in branches.

We identified several significant associations between age and mutation timing. In pan-PCAWG analysis, we found positive associations between age and proportion of clonal SNVs (ρ = 0.20, adjusted LNR p = 1.4×10^{-3}, **Figure 1C**) and proportion of clonal indels (ρ = 0.14, LNR p = 0.013, **Supplementary Table 2**). Age was also associated with increasing clonal SNV proportion in two tumour-types: stomach cancer (Stomach-AdenoCA: ρ =0.44, adjusted LNR p = 0.028), and medulloblastoma (CNS-Medullo: ρ = 0.34, adjusted LNR p = 2.5×10^{-3}, **Figure 1C**). Positive correlations in these tumour-types indicate tumours arising in older individuals accumulated a greater fraction of SNVs earlier in tumour evolution. In contrast, we identified the inverse trend in melanoma, where tumours of younger patients accumulated more subclonal than clonal SNVs (ρ = -0.47, adjusted LNR p = 7.8×10^{-3}). Differences in the proportion of clonal mutations suggest differential mutation timing over the tumour evolution and could be caused by mechanisms such as differences in mutational processes or driver mutation frequency.

We next focused on genome instability, a measure of copy number aberration (CNA) burden and approximated by the percent of the genome altered by CNAs (PGA). Analogous to SNV density measuring the burden of point mutations, PGA measures the density of copy number alterations. We found that in pan-cancer analysis, PGA increased with age in PCAWG (ρ = 0.19, adjusted LNR p = 0.022) and AACR GENIE (ρ = 0.041, adjusted LNR p = 0.050) (**Figure 1D**) and estimate that PGA increases at 0.010% per year (**Table 2**). We also identified positive correlations in six TCGA, three PCAWG, and three AACR GENIE tumour-types. Again, prostate cancer showed consistent age-PGA associations, this time in all three datasets (TCGA: ρ = 0.17, adjusted LNR p = 6.7×10^{-5}; PCAWG: ρ = 0.27, adjusted LNR p = 3.0×10^{-3}; AACR GENIE: ρ = 0.11, adjusted LNR p

= 0.050; **Figure 1E**). Age was associated with PGA in stomach cancer data in TCGA (ρ = 0.11, adjusted LNR p = 0.011) and AACR GENIE (ρ = 0.38, adjusted LNR p = 0.041), and while other age-PGA correlations were not statistically significant across multiple datasets, they showed similar effect sizes (**Supplementary Figure 1**). Intriguingly, we detected negative age-PGA associations in TCGA lung adenocarcinomas (**Figure 1D**), and correspondingly negative associations in PCAWG (ρ = -0.13) and AACR GENIE lung tumours (ρ = -0.099) (**Supplementary Figure 1**). Estimates for per year increase in PGA for these tumour-types are given in **Table 2** for tumour-types with consistent evidence in at least two datasets.

To better understand why lung cancers of younger patients might have higher genomic instability, we investigated tobacco use. Tobacco exposure is a well-known risk factor in lung cancers. Our multivariate lung cancer regression models do account for tobacco use, but we supplemented our analysis with a more focused study of the relationship between PGA, age and tobacco history. Age was associated with tobacco history (Kruskal-Wallis $p = 1.6 \times 10^{-12}$, **Figure 1F, top**), where current smokers were diagnosed with lung cancer at younger ages than never and reformed smokers (difference in location = -6.0, 95% CI = -8.0 – -4.0, Wilcoxon rank sum test $p = 1.2 \times 10^{-7}$). PGA was also associated with smoking history (Kruskal-Wallis $p = 1.6 \times 10^{-12}$, **Figure 1F, bottom**) where tumours arising in current smokers had higher PGA (difference in location = 5.0, 95% CI = 1.2 – 9.2, Wilcoxon rank sum test $p = 1.0 \times 10^{-3}$). We then examined the associations between age and PGA within each tobacco use category and found no association in never and reformed smokers ≥ 15 years (Spearman correlation p ≥ 0.1, **Figure 1G**). In contrast, tumours of current and reformed smokers < 15 showed statistically significant negative correlations, indicating the negative association between age and PGA are dependent on current or recent tobacco use. These mutation density results demonstrate nuances in the relationship between aging and the accumulation of mutations: there are differences by when mutations occur throughout tumour evolution, and by how they occur through different mutational processes.

Age-Associated Mutational Signatures

We continued exploring mutational processes through mutational signatures data. Clinical data on exposures such as tobacco use history and frequency are self-reported by the patient and may be inaccurate, mis-recorded, or not described. Moreover, clinical data often do not include information on exposures such as second-hand smoke, which is also a known cancer risk factor[59]. We leveraged the COSMIC mutational signatures, which can be applied to deconvolve distinctive mutational patterns from genomic sequencing[60]. Each mutational signature is thought to represent a specific oncogenic process – for instance, single base signature (SBS) 4 represents tobacco exposure. Using mutational signatures, we can not only detect the influence of oncogenic processes within a tumour, but we can also quantify the number of mutations attributed to that

process and assess its relative activity compared with the activity of other mutational processes active within the tumour.

Focusing on the tobacco smoking signature SBS4, we stratified TCGA lung adenocarcinoma tumours into SBS4-postive (SBS4+) and -negative (SBS4-) groups. SBS4+ tumours were detected more frequently in younger patients (Wilcox $p = 1.1x10^{-5}$) and have higher PGA (Wilcox $p =1.0x10^{-3}$, **Figure 2A left**). We also found that the number of mutations attributed to SBS4 was negatively correlated with age (Spearman's $\rho = -0.12$, p-value = 0.086) and positively correlated with PGA (Spearman's $\rho = 0.31$, p-value $= 2.9x10^{-6}$, **Figure 2A centre**). These findings show that a greater burden of tobacco-associated mutations occurs in lung tumours of younger patients and is associated with increased genomic instability. When considering SBS4-attributed SNVs as a fraction of total detected SNVs, we found that while age remained negatively correlated with SBS4 relative activity (Spearman's $\rho = -0.16$, p-value = 0.017), there was no association with PGA (Spearman's $\rho = 0.082$, p-value = 0.23, **Figure 2A right**): while smoking-attributed mutations account for a greater percentage of mutations in younger patients, this relative activity is not associated with a change genomic instability. Also, when examining the relationships between age and PGA separately in SBS4+ and SBS4- tumours, we found significant negative correlations in both groups (**Figure 2B**), suggesting additional factors beyond SBS4-attributed mutations may contribute to increased PGA in lung tumours diagnosed in younger patients.

We repeated this analysis in TCGA lung squamous cell cancers and detected a similar negative correlation between age and PGA in SBS4- tumours (Spearman's $\rho = -0.17$, p-value $= 5.8x10^{-3}$, but no association in SBS+ tumours (Spearman's $\rho = -0.0069$, p-value = 0.94, **Supplementary Figure 2**). SBS4 activity was also negatively associated with age in PCAWG lung adenocarcinomas (Lung-AdenoCA: $\rho = -0.50$, adjusted LNR p = 0.025, **Supplementary Figure 2**). Indeed, SBS4 and age were consistently negatively associated across both subtypes of lung cancer and both datasets, though not all associations were statistically significant after multiple testing adjustment. This supports previous findings that tobacco has a larger tumorigenesis role in younger patients, with tobacco-associated mutations contributing to a greater portion of the mutational landscape of tumours derived from younger individuals[61].

The PCAWG project updated the COSMIC signatures v3 to a set of 49 single base substitution , 11 doublet base substitution (DBS) and 17 small insertion and deletion (ID) signatures[62]. We extended our analysis of SBS4 to all 77 mutational signatures in PCAWG data and to SBS signatures in TCGA data. We did not investigate mutational signatures in AACR GENIE data due to smaller mutation numbers resulting from the limited genome coverage of the MSK-IMPACT panel. Similar to our analysis of SBS4, we examined both the proportion of signature-positive tumours as well as relative mutation activity. Previous studies of mutational signatures describe the correlations between age

and signature-attributed mutations but ignore the other aspects of signature detection and relative activity. By comparing signature detection rates, we identify mutational processes that are more likely to be active in younger *vs* older patients and vice versa. By analyzing signature-attributed mutations as a proportion of total mutations per tumour, we can derive information about that signature's contribution to the overall mutational spectrum. For example, SBS1 is well-known as being 'clock-like' and its number of attributed mutations increase with age[60,62]. However, because SBS1 is detected almost universally, it is equally likely to occur in tumours of younger *vs* older patients; when analysed as a proportion of total mutations, we find that the proportion of SBS1 mutations does not change with age, suggesting that its relative activity is stable with age (Spearman's correlation $p > 0.1$).

Across all 2,562 PCAWG tumours, we identified twelve mutational signatures with age-associated detection frequency (**Figure 2C, left**) and ten with age-associated relative signature activity (**Figure 2D, left**). For example, tumours arising in older patients were more likely to be SBS3-positive (marginal log odds change = 0.0085, 95%CI = 0.0024-0.015, adjusted LGR $p = 0.075$), but in these SBS-positive tumours, the proportion of SBS3-attributed mutations decreased with age ($\rho = -0.20$, adjusted LNR $p = 3.2 \times 10^{-3}$). SBS3 mutations are thought to be caused by defective homologous recombination-based DNA damage repair. These results imply that while tumours derived from older individuals are more likely to harbour defective DNA damage repair, its relative impact on the burden of SNVs is lower compared with tumours derived of younger individuals. A similar relationship was seen for ID8, which is linked to defective non-homologous DNA end-joining (marginal log odds change = 0.024, 95%CI = 0.020-0.028, adjusted LGR $p = 3.4 \times 10^{-3}$; $\rho = -0.099$, adjusted LNR $p = 3.7 \times 10^{-5}$) and ID1, associated with slippage during DNA replication (marginal log odds change = 0.013, 95%CI = 0.0059-0.020, adjusted LGR $p = 0.018$; $\rho = -0.059$, adjusted LNR $p = 0.048$). Like our results for SBS4, we identified associations between age and other tobacco-related signatures DBS2 and ID3. Conversely, tumours arising in older individuals were less likely to exhibit defective base excision repair (SBS36). All mutations signatures findings are in **Supplementary Table 2**.

These pan-cancer differences persisted across individual tumour-types. We identified 23 age-associated signatures across eleven tumour-types, including six significant signatures in melanoma. In this tumour-type, tumours arising in older patients were preferentially SBS2-positive (marginal log odds change = 0.051, 95%CI = 0.013-0.095, adjusted LGR $p = 0.029$, **Figure 2C**), attributed to APOBEC cytidine deaminase activity[63]. Melanomas arising in younger patients were more likely to be positive for signatures related to UV damage (SBS 7a, 7b, 7d, **Figure 2C, Supplementary Table 2**). The proportion of mutations attributed to UV damage was also higher in younger patients (DBS1, $\rho = -0.29$, adjusted LNR $p = 0.019$, **Figure 2D**), while the proportion of mutations attributed to slippage during DNA replication was higher in older patients (ID1, $\rho = 0.27$,

adjusted LNR p = 0.019, **Figure 2D**). These results suggest that melanomas in younger patients more frequently involve UV exposure and damage, while melanomas in older patients are more influenced by endogenous sources of mutation.

Leveraging data describing SBS signatures in TCGA data, we repeated this analysis to identify age-associations in signatures derived from whole exome sequencing (WXS) data. Across pan-TCGA tumours, we detected five signatures that occurred more frequently in older individuals, and three that occurred more frequently in younger individuals (**Figure 2C**). We also identified five signatures with higher relative activity in younger patients (**Figure 2D**).

There was moderate agreement between TCGA and PCAWG findings: while the results in one dataset never contradicted those of the other, some signatures were associated with age exclusively in either TCGA or PCAWG data. Other signatures, such as SBS1 and SBS5 were associated with age in detection and relative activity across a range of tumour-types in either dataset. There was complete agreement in only some signature like SBS2 and SBS4. We hypothesized that this was due to differences in signature detection rates between WXS and whole genome sequencing (WGS) data and compared how frequently each signature was detected across all samples (**Figure 2E**). Signatures with high agreement between datasets had similar detection rates, as observed for SBS2 (detection difference = 1.5%) and SBS4 (detection difference = 1.1%). Signatures where findings did not replicate had vastly different detection rates, as was seen for SBS1 (detection difference = 7.2%) and SBS5 (detection difference = 10%). We further examined this by comparing signatures data from non-PCAWG WGS and non-TCGA WXS data. Differences in signature detection rates between PCAWG and TCGA data were reflected in non-PCAWG WGS and non-TCGA WXS data (**Supplementary Figure 2**). We also looked specifically at identified age-associations and found high agreement in data generated by the same sequencing strategy (**Supplementary Figure 2**). These findings suggest high confidence in age-associations detected in both WGS and WXS data, and that additional study is needed in independent WXS and WGS data to validate TCGA- and PCAWG-specific findings.

CNA Differences Associated with Transcriptomic Changes

Global mutation characteristics such as genome instability are features of later stages in a tumour's evolutionary history. In contrast, the early stages are often driven by chromosome- or gene-specific events such as loss of specific chromosomes or mutation of driver genes[55]. We therefore narrowed our focus to 87 cancer driver genes frequently affected by copy number alterations as described by COSMIC[64]. We applied our statistical framework to identify putative age-associated copy number driver gains and losses using univariate logistic regression. Putative events identified with a false discovery rate threshold of 10% were further analysed by multivariable logistic regression to account for confounding factors. We applied these analyses to PCAWG, TCGA, and AACR GENIE

datasets separately to characterize pan-cancer and tumour-type-specific associations. We assessed our statistical approach with p-value Q-Q plots as presented in **Supplementary Figure 4**.

In pan-cancer analysis of TCGA data, we identified 23 driver genes that were more frequently lost and eight driver genes that were more frequently gained in in tumours of older individuals (**Figure 3A**). Age-associated loss of *FANCA* (marginal log odds change = 0.015, 95%CI = 0.012-0.018, adjusted MLR p = 3.2×10^{-9}) was also statistically significant in AACR GENIE data (marginal log odds change = 0.051, 95%CI = 0.036-0.066, adjusted MLR p = 0.011). Other age-associated loss and gain events were statistically significant in one dataset and corroborated by similar effects in at least one other. For example, loss of *STK11* was significant in PCAWG (marginal log odds change = 0.030, 95%CI = 0.023-0.037, adjusted MLR p = 0.014), and was positively associated with age in TCGA (marginal log odds change = 0.014) and AACR GENIE (marginal log odds change = 0.044) datasets without reaching our statistical significance thresholds (**Supplementary Table 3-4**).

There were also age-associated CNAs in specific tumour-types: we detected age-associated gains with evidence in at least two datasets in eight tumour-types (**Figure 3B**), and losses in ten tumour-types (**Figure 3C**). Most of these associations were positive, indicating they were more likely to occur in tumours of older patients. We found the most evidence of age-associated drivers gains in ovarian cancer across all three datasets, such as gain of *ALK* (marginal log odds changes: TCGA = 0.016; PCAWG = 0.070; AACR GENIE = 0.071) and MDM4 (marginal log odds changes: TCGA = 0.023; PCAWG = 0.056; AACR GENIE = 0.071; **Supplementary Table 4**). There were also eight age-associated driver losses in this tumour-type. All identified CNAs were more frequently detected in tumours of older patients except for *TNFAIP3* loss, which occurred more frequently in tumours of younger patients (marginal log odds changes: TCGA = -0.028; PCAWG = -0.071; **Supplementary Table 3**). Other CNAs more likely to occur in tumours of younger patients were detected in colorectal, lung, breast, colorectal, liver, glioblastoma, and melanoma tumour-types (**Figure 3B-C**, **Supplementary Tables 3-4**).

We next asked whether age-associated CNAs might lead to downstream transcriptomic changes by investigating TCGA tumour-matched mRNA abundance data. We used linear models with age, copy number status, and the interaction between age and copy number status as predictors. These terms tell us when mRNA abundance differs by age, when the CNA event itself is significantly associated with mRNA abundance, and when the effect of the CNA event on mRNA depends on age. We also adjusted for tumour purity (as estimated by study pathologists) in all mRNA analyses. Of 58 age-associated CNAs with mRNA data, we found 17 were significantly associated with changes in mRNA abundance (**Figure 3D left, Supplementary Table 5**). For example, loss of *CDKN2A* in glioblastoma, more frequent in tumours of older individuals, was associated with

decreased mRNA abundance (adjusted mRNA-CNA p = 1.3×10^{-3}). Intriguingly, *CDKN2A* loss in sarcoma was not significantly associated with a change in mRNA on its own but did have a significant interaction with age (**Figure 3D right**). There was a larger decrease in *CDKN2A* mRNA in sarcomas derived of older individuals than younger individuals (adjusted mRNA-CNA-age p = 0.024, **Figure 3E**). Thus, these age-associated CNAs not only preferentially occur in tumours derived of individuals of a certain age group, they are also associated with changes in mRNA abundance.

To investigate potential clinical significance of these age-associated CNAs, we performed survival analysis to identify prognostic events. We used Cox Proportional-Hazards (Cox PH) models with 5-year overall survival as the end point. Like our mRNA models, we used predictors including copy number status, age and their interaction. In glioblastoma, age itself is a known prognostic feature with older patients having poorer outcome (HR = 2.1, 95%CI = 1.7-2.6, Wald p = 1.4×10^{-13}). We found that loss of *SUFU* is itself not prognostic, but incorporating age revealed that younger individuals with no *SUFU* loss have the best outcome (HR = 0.42, 95%CI = 0.30 - 2.3, adjusted Wald p = 5.5×10^{-6}). Also, *SUFU* loss stratifies younger individuals into two groups with distinct trajectories, but has no such prognostic value in glioblastomas of older individuals (**Figure 3F**).

We repeated these mRNA and survival analyses for all TCGA tumour-types with age-associated CNAs. Genes in these CNAs were associated with altered mRNA abundance in six tumour-types. Five tumour-types also showed that age-associated CNAs can be prognostic and that the prognostic value can also differ based on the age of the individual. We were unable to repeat these analyses in PCAWG data due the small number of patients with mRNA abundance and outcome data but present all mRNA and survival analysis results in **Supplementary Table 5**.

SNVs Differences Associated with Functional Changes

Finally, we investigated gene-level SNVs for age-associations. In PCAWG analysis, we used a predefined set of genomic driver and mitochondrial genes[65]. In TCGA analysis, we focused on a set of 679 COSMIC driver genes[64] and applied a recurrence threshold to filter out genes mutated in less than 1% of tumours. We used AACR GENIE data generated using MSK-IMPACT targeted sequencing of up to 468 cancer genes, and we applied a recurrence threshold of 1%. We included SNV density in our multivariate models in addition to other confounding factors as previously described.

In pan-cancer analysis, we identified 102 age-associated genes in TCGA, nine age-associated genes in AACR GENIE, and one in PCAWG data (**Supplementary Table 6**). *CREBBP*-frequency was associated with age in both TCGA (marginal log odds change = 0.030, 95%CI = 0.024 – 0.040, adjusted LGR p = 0.049) and PCAWG (marginal log odds change = 0.027, 95%CI = 0.0089 – 0.047, adjusted LGR p = 8.7×10^{-3}, **Figure 4A, Supplementary Table 6**). In AACR GENIE, the positive association between *CREBBP*-status and age was not significant after multiple testing correction (marginal log odds

change = 0.011, 95%CI = 0.0047 – 0.022, adjusted p > 0.1). *KDM6A* and *RBM10* were more likely to have SNV in tumours derived of older patients in TCGA and AACR GENIE but were not recurrently mutated in PCAWG data and not analysed in that dataset (**Figure 4A, Supplementary Table 6**). 35 other genes found to be associated with age in TCGA data showed similar effect sizes in either PACWG or AACR GENIE data, but without reaching statistical significance.

There were also tumour-type specific age-associations in SNV frequency in TCGA, PCAWG and AACR GENIE. We identified three tumour-types with consistent and significant age-associated SNVs in at least two datasets, and five tumour-types with associations that were significant in one dataset and showed the same effect in one other (**Figure 4B, Supplementary Table 6**). SNVs in *FOXA1* occurred more frequently in breast and prostate tumours derived of older individuals. *SPOP* was also positively associated with age in prostate tumours (PCAWG adjusted LGR p = 0.099, AACR GENIE adjusted LGR p = 0.03). In melanoma, SNVs in *NF1* were more frequent in tumours of older individuals, while *BRAF* SNVs were more frequent in tumours of younger individuals. We also confirmed known associations between age and mutations in tumour suppressors *IDH1-* and *ATRX-* in both high grade glioblastoma and lower grade gliomas, *IHD1-* and *ATRX-* were more frequent in tumours derived of younger individuals (**Figure 4C**, **Supplementary Table 6**). Other age-associated SNVs include positive associations in cervical and head and neck cancer, and negative associations in colorectal cancer (**Supplementary Table 6**).

Like the nuclear genome, the mitochondrial genome is frequently mutated in cancer[66]. We leveraged mitochondrial SNV (mtSNV) data from PCAWG whole genome sequencing and identified age associated mtSNVs in pan-cancer analysis and in ovarian cancer (**Figure 4D**, **Supplementary Table 6**). All significant age-associations were mtSNVs that occurred more frequently in tumours of older patients, even after controlling for the number of mitochondria copies in each tumour. Implicated mitochondrial regions include *MT-CYB,* which encodes a cytochrome b and the D-loop, a non-coding region that controls replication and transcription[67,68] (**Figure 4E)**. We also examined whether the number of tumour mitochondria copies was associated with age by investigating the foldchange compared with normal mitochondria copies. There was indeed a significant association between mitochondria copy increase and age in pan-cancer analysis and three tumour-types (**Figure 4F**, **Supplementary Table 6**). In these tumour contexts, tumours of older patients gained more mitochondria copies than in tumours of younger patients.

As with the age-associated CNAs, we evaluated the impact of SNVs on mRNA abundance and survival in TCGA data. We identified significant associations between age-associated SNVs and mRNA abundance for *ATRX* and *IDH1* in lower grade glioma (**Supplementary Table 5**). Mutations in *ATRX* and *IDH1* were associated with lowered

mRNA abundance in both genes. There was also a significant interaction between age and *IDH1*- frequency (adjusted $p = 2.1 \times 10^{-4}$, **Figure 4G**) indicating an age-dependent effect on mRNA abundance: mutated *IDH1* was associated with a greater mRNA decrease in tumours arising in younger patients. Interestingly, this difference results from a change in baseline *IDH1* mRNA: older patients have higher *IDH1* mRNA abundance than younger, and mutated *IDH1* leads to equalised mRNA levels. *IDH1* encodes isocitrate dehydrogenase 1, a component of the citric acid cycle: differences in its baseline abundance may be due to differences in metabolism in younger and older brains[69]. In contrast while there was no interaction between age and mutation status on mRNA abundance (adjusted $p = 0.36$, **Figure 4G**), *ATRX*- and age were synergistically associated with outcome, stratifying lower grade glioma patients into four groups (**Figure 4H**).

The lower grade glioma tumour-type is comprised of astrocytoma and oligodendroglioma subtypes. *IDH1* mutation is intrinsically linked to glioma subtype as oligodendroglioma is diagnosed based on the presence of both the 1p/19q co-deletion and mutation of either *IDH1* or *IDH2*[70]. While our multivariable models adjust for tumour subtype, we investigated age-associations of *ATRX* and *IDH1* SNV mutation frequency in lower grade gliomas in greater detail. We stratified the TCGA lower grade tumours into astrocytoma, oligodendroglioma and oligoastrocytoma subtypes and repeated the SNV, mRNA and survival analyses in each group. We found *TP53* and *ATRX* SNV mutations occurred more frequently in tumours derived of younger individuals across all three subtypes (Wilcox **Supplementary Figure 5**). *ATRX* SNV mutation frequency was associated with decreased mRNA abundance in all three subtypes, and we also found a significant age-*ATRX* interaction in astrocytoma, where *ATRX*- was associated with a greater decreased in mRNA abundance in tumours of older individuals than younger (interaction $p = 0.016$, **Supplementary Figure 5**). *IDH1* SNV mutation frequency was negatively associated with age in astrocytoma (Wilcox $p = 1.3 \times 10^{-8}$) and oligoastrocytoma (Wilcox $p = 0.041$), but not oligodendroglioma (Wilcox $p = 0.19$), and we found significant age-*IDH1* interactions in mRNA abundance analysis for oligodendroglioma (interaction $p = 0.038$) and oligoastrocytoma (interaction $p = 6.2 \times 10^{-3}$). Finally, we found significant age-dependent associations of *ATRX* SNV status in astrocytoma but not oligodendroglioma or oligoastrocytoma: ATRX SNVs were associated with improved survival in older patients, but with worse survival in younger patients.

Discussion

Despite modest statistical power, suboptimal study designs and limited clinical annotation, we identified myriad age-associated differences in cancer genomes. Age-associated genomic features occur at the pan-cancer level and also across almost all individual tumour-types. Combined with similar reports of sex- and ancestry-associated differences in cancer genomes[40,71], these data reveal a set of host influences on the mutational characteristics of tumours. Indeed, a study in simultaneous preparation by Chatsirisupachai *et al.* describes corroborating evidence of age-associated differences in the genome and transcriptome, as well as age-specific differences in methylation and gene expression control[72]. Together, we find that characteristics of the tumour host appears to influence all aspects of the tumour molecular profile and that some of these lead to age-specific transcriptomic and clinical impacts. We note that throughout this study we report Bonferroni-adjusted p-values, rather than FDR-adjustment which is standard in our field. This change was required in peer-review, and provides our results significantly more statistical stringency that than common in genomic studies.

The mechanisms for these genomic associations are largely unknown. Our data suggest some endogenous or exogenous mutational processes preferentially occur in individuals of different age groups. Some of these mutational processes are related to aging-associated phenomena such as declining DNA damage repair[73,74], somatic mosaicism and the accumulation of mutations over time[60,75,76]. However, other processes related to immune surveillance, evolutionary selection, disease aetiology and epigenetics are also likely involved[77–79]. Pathogenic germline variants such as those in *BRCA1/2* or *TP53* also lead to earlier presentation of cancer. While our results remain unchanged on removing tumours with detected known pathogenic variants[80], it is likely there remains hereditary confounders that we have not accounted for.

In addition to such biological factors, lifestyle and socioeconomic considerations like diet[81] and microbiome composition[82] can continuously shape tumour evolution from its earlier steps. Many of these factors are deeply linked to not only an individual's age, but other fundamental characteristics over which we have limited control, such as ancestry or sex. For example, we found that tobacco exposure is closely linked to the negative correlation between age and PGA. It is possible that tobacco exposure leads to earlier presentation of mutation-dense lung cancers. However, it is also likely that there are other variables and interactions that influence the relationship between age and mutation density. Moreover cohort effects, where individuals born in one time period experience different risk exposures from those born in another, can greatly influence the somatic profile of tumours. Our analyses do not consider such cohort effects, and some described age-associations may instead be attributed to differences across time. A tumour's mutational history therefore reflects a complex interplay of biological, lifestyle and healthcare factors,

and we have little understanding of how these diverse processes interact to produce molecular phenotypes.

Most tumour-types in our study showed some association between age and genomic alterations. Prostate cancer showed persistent associations in all three datasets across multiple measures of genomic alterations. These were all positive age-associations where mutation density, driver CNA frequency, and SNV frequency of *SPOP* and *FOXA1* all increased with age. The strong association of age with alterations in prostate cancer may be related to its typically slow-growth and low mutation burden. Moreover, known exogenous risk factors, including diet and endocrine disrupting chemicals are thought to converge on hormone regulation, which then acts on the prostate[83]. In contrast, tumours types with negative age-associations such as lung and liver cancers have exogenous risk factors that impact cells through multiple indirect and direct mechanisms such as viral infection, mutagenesis and inflammation[84].

The TCGA, PCAWG and AACR GENIE datasets sometimes identified different molecular associations, highlighting the differences between the three datasets. TCGA and AACR GENIE patients were largely North American while PCAWG had a greater international component. While the ages represented in all three datasets were similar (**Table 1**), the cohorts differ in other host and clinical characteristics. For instance, the representation of ancestry groups is dissimilar, with many tumour-types differing vastly in ancestry proportions (**Supplementary Table 1**). Furthermore, differences in sequencing targets also contributed to variation in our results, most conspicuously in the detection rates of some mutational signatures. We customized our analyses to take advantage of the contrasting strengths of each dataset: WGS in PCAWG allowed us to interrogate a greater breadth of mutation types, while the larger sample size and clinical annotation of TCGA data improved statistical power and controls for confounders. The focused panel used to generate AACR GENIE data facilitated the validation of age-associations in cancer genes. Also, while we were able to identify more age-associations in TCGA data, many of these findings were reflected in PCAWG and AACR GENIE data by similar effects that did not reach our statistical significance threshold. More sequencing data reflecting greater and more balanced diversity is needed to distinguish those age-associations that are intrinsic to differences in biology, and those that are tied to differences in lifestyle and geography.

Our findings have wide-reaching implications for both basic and translational cancer research. Since cancer host characteristics like age, ancestry and sex widely shape the somatic cancer landscape, we cannot consider discovery genomics complete they are explicitly considered. Elderly individuals are underrepresented in cancer sequencing studies and clinical trials[39,85,86]: better inclusion is needed to identify somatic changes specific to older individuals and to leverage these changes to improve clinical care. In our analysis, we found that some age-associated genomic differences associate with

transcriptional and clinical changes, but many do not – identifying the functional consequences and mechanisms of these will be a long-term challenge. Finally, these epidemiological factors should be considered and controlled for in personalized therapy strategies. Indeed, every type of analysis from driver-discovery to biomarker-development should explicitly test for and model the powerful influence of patient biology and behaviour on tumour evolution.

Online Methods

Data acquisition & Processing

Genome-wide somatic copy-number, somatic mutation, and mRNA abundance profiles for the Cancer Genome Atlas (TCGA) datasets were downloaded from Broad GDAC Firehose (https://gdac.broadinstitute.org/), release 2016-01-28. For mRNA abundance, Illumina HiSeq rnaseqv2 level 3 RSEM normalised profiles were used. Genes with >75% of tumours having zero reads were removed from the respective dataset. GISTIC v2 (13) level 4 data was used for somatic copy-number analysis. mRNA abundance data were converted to $\log_2$ scale for subsequent analyses. Mutational profiles were based on TCGA-reported MutSig v2.0 calls. All pre-processing was performed in R statistical environment (v3.1.3). Genetic ancestry imputed by Carrot-Zhang[87] *et al*. was incorporated.

PCAWG whole genome sequencing data calls were downloaded from the PCAWG consortium through Synapse. All data pre-processing was performed by the consortium as described[88]. The individual datasets are available at Synapse (https://www.synapse.org/) and are denoted with synXXXXX accession numbers (listed under Synapse ID); all these datasets are also mirrored at https://dcc.icgc.org. Tumour histological classifications were reviewed and assigned by the PCAWG Pathology and Clinical Correlates Working Group (annotation version 9; https://www.synapse.org/#!Synapse:syn10389158, https://www.synapse.org/#!Synapse:syn10389164). Ancestry imputation was performed using an ADMIXTURE23-like algorithm based on germline SNP profiles determined by whole-genome sequencing of the reference sample (https://www.synapse.org/#!Synapse:syn4877977). The consensus somatic SNV and indel (https://www.synapse.org/#!Synapse:syn7357330) file covers 2778 whitelisted samples from 2583 donors. Driver events were called by the PCAWG Drivers and Functional Interpretation Group (https://www.synapse.org/#!Synapse:syn11639581). Consensus CNA calls from the PCAWG Structural Variation Working Group were downloaded in VCF format (https://www.synapse.org/#!Synapse:syn8042988). Subclonal reconstruction was performed by the PCAWG Evolution and Heterogeneity Working Group (https://www.synapse.org/#!Synapse:syn8532460). SigProfiler mutation signatures were determined by the PCAWG Mutation Signatures and Processes Working Group for single base substitution (https://www.synapse.org/#!Synapse:syn11738669), doublet base substitution (https://www.synapse.org/#!Synapse:syn11738667) and indel (https://www.synapse.org/#!Synapse:syn11738668) signatures. Signatures data for TCGA, non-PCAWG WGS and non-TCGA WXS samples were downloaded from Synapse (https://www.synapse.org/#!Synapse:syn11804040).

Controlled access AACR GENIE data was downloaded from AACR Project GENIE for the MSK project[89]. (https://portal.gdc.cancer.gov/projects/GENIE-MSK). AACR GENIE MSK data was generated by the custom hybridization-based capture panel MSK-IMPACT, which includes up to 410 genes. Preprocessed gene-wise SNV and CNA calls were obtained through the GDC Data Portal. We filtered 6,841 metastatic tumours and matched on tumour-types included in the TCGA and PCAWG datasets (**Supplementary Table 1**) for better comparison across datasets. The final pan-cancer AACR GENIE cohort consisted of 7,259 tumours across 35 tumour-types.

We used TCGA data describing 10,212 distinct TCGA tumour samples across 23 tumour-types and 2,562 distinct PCAWG samples across 29 tumour-types. Tumour-types with no age annotation were excluded from analysis. Age is treated as a continuous variable for both TCGA and PCAWG analyses. We matched tumour-types between datasets as described in **Supplementary Table 1**.

General Statistical Framework

For each genomic feature of interest, we used univariate tests first followed by false discovery rate (FDR) adjustment to identify putative age-associations of interest ($q < 0.1$). We used two-sided non-parametric univariate tests to minimize assumptions on the data. For putative age-associations, we then follow up the univariate analysis with multivariate modeling to account for potential confounders using bespoke models for each tumour-type.

Model variables for each tumour context are presented in **Supplementary Table 1** and were included based on availability of data (<15% missing), sufficient variability (at least two levels) and collinearity (as assessed by variance inflation factor). Discrete data was modeled using logistic regression (LGR). Continuous data was first transformed using the Box-Cox family and modeled using linear regression (LNR). The Box-Cox family of transformations is a formalized method to select a power transformation to better approximate a normal-like distribution and stabilize variance. We used the Yeo-Johnson extension to the Box-Cox transformation that allows for zeros and negative values[90].

FDR adjustment was performed for p-values for the age variable significance estimate and an FDR threshold of 10% was used to determine statistical significance. More detail is provided for each analysis below. A summary of all results is presented in **Supplementary Table 1**. We present 95% confidence intervals for all tests.

Mutation Density

Performed for both TCGA, PCAWG and AACR GENIE data. Overall mutation prevalence per patient was calculated as the sum of SNVs across all genes on the autosomes and scaled to mutations/Mbp. Coding mutation prevalence only considers the coding regions of the genome, and noncoding prevalence only considers the noncoding regions. TCGA mutation density reflects coding mutation prevalence. AACR GENIE mutation density

reflects the targeted sequences of the 341-, 410- or 468-gene MSK-IMPACT panel used for each sample. Mutation density was compared age using Spearman correlation for both pan-cancer and tumour-type specific analysis. Comparisons with univariate q-values meeting an FDR threshold of 10% were further analyzed using linear regression to adjust for tumour subtype-specific variables. Mutation density analysis was performed separately for each mutation context, with pan-cancer and tumour subtype p-values adjusted together. Full mutation density results are in **Supplementary Table 2.**

Per year increase in SNV density was estimated by combining TCGA, PCAWG and AACR GENIE data. We fit a linear regression model with formula SNV density ~ age + project and took the coefficient estimate for age as the per year increase in SNV density value.

Genome instability

Performed for TCGA, PCAWG and AACR GENIE data. Genome instability was calculated as the percentage of the genome affected by copy number alterations. The number of base pairs for each CNA segment was summed to obtain a total number of base pairs altered per patient. The total number of base pairs was divided by the number of assayed bases to obtain the percentage of the genome altered (PGA). Genome instability was compared using Spearman correlation for both pan-cancer and tumour-type specific analysis. Comparisons with univariate q-values meeting an FDR threshold of 10% were further analyzed using linear regression to adjust for tumour subtype-specific variables. Genome instability analysis was performed separately for each mutation context, with pan-cancer and tumour subtype p-values adjusted together. Full mutation density results are in **Supplementary Table 2.** Per year increase in PGA was estimated similarly to the estimation for SNV density: we fit a linear regression model with formula PGA ~ age + project and took the coefficient estimate for age as the per year increase in PGA.

Clonal structure and mutation timing analysis

Performed for PCAGW data only. Subclonal structure data was binarized from number of subclonal clusters per tumour to monoclonal (one cluster) or polyclonal (more than one cluster). Putative age-associations were identified using univariate logistic regression and putative associations were further analysed using multivariate logistic regression. A multivariate q-value threshold of 0.1 was used to determine statistically significant age-associated clonal structure.

Mutation timing data classified SNVs, indels and SVs into clonal (truncal) or subclonal groups. The proportion of truncal variants was calculated for each mutation type ($\frac{Number\ truncal\ SNVs}{total\ SNVs}$, *etc.*) to obtain proportions of truncal SNVs, indels and SVs for each tumour. These proportions were compared using Spearman correlation. Univariate p-values were FDR adjusted to identify putatively age-associated mutation timing. Linear regression was used to adjust for confounding factors and a multivariate q-value threshold

of 0.1 was used to determine statistically significant age-a mutation timing. The mutation timing analysis was performed separately for SNVs, indels and SVs. All results for clonal structure and mutation timing analyses are in **Supplementary Table 2**.

Mutational Signatures analysis

Performed for TCGA and PCAWG data. For each signature, we compared the proportion of tumours with any mutations attributed to the signatures ("signature-positive") using logistic regression to identify univariate age-associations. Signatures with putative age-associations were further analysed using multivariable logistic regression.

We also compared relative signature activity using the proportions of mutations attributed to each signature. The numbers of mutations per signature were divided by total number of mutations for each tumour to obtain the proportion of mutations attributed to the signature. Spearman correlation was used. Putative age-associated signatures were further analysed using multivariable linear regression after Box-cox adjustment.

Signatures that were not detected in a tumour subtype was omitted from analysis for that tumour subtype. All results for clonal structure and mutation timing analyses are in **Supplementary Table 2.**

Genome-spanning CNA analysis

Performed for TCGA, PCAWG and AACR GENIE data. The copy number profiles for COSMIC driver CNA genes were extracted. Copy number calls were collapsed to ternary (loss, neutral, gain) representation by combining loss groups (mono-allelic and bi-allelic) and gain groups (low and high). Logistic regression was used to identify univariate age-associated CNAs. After identifying candidate pan-cancer univariately significant genes, multivariate logistic regression was used to adjust ternary CNA data for tumour-type-specific variables. The genome-spanning analysis was performed separately for losses and gains for each tumour subtype. All CNA results are in **Supplementary Tables 3-4**.

Driver SNV Analysis

Performed for TCGA, PCAWG and AACR GENIE data. We focused on driver events described by the PCAWG consortium[65] and by COSMIC[64]. For TCGA and AACR GENIE, we also applied a 1% recurrence filter. Driver mutation data was binarized to indicate presence or absence of the driver event in each patient. Proportions of mutated genes were compared using univariate logistic regression. A q-value threshold of 0.1 was used to select genes for further multivariate analysis using binary logistic regression. SNV density was included in all models. FDR correction was again applied and genes with significant age terms were extracted from the models (q-value < 0.1). Driver event analysis was performed separately for pan-cancer analysis and for each tumour subtype. All SNV and driver event analysis results are in **Supplementary Table 6**.

mRNA functional analysis

Performed for TCGA data. Genes in bins altered by age-associated CNAs and SNVs after multivariate adjustment were further investigated to determine functional consequences. Tumour purity was included in all mRNA models. Tumours with available mRNA abundance data were matched to those used in CNA analysis. For each gene affected by an age-associated loss, its mRNA abundance was modeled against age, copy number loss status, an age-copy number loss interaction term, and tumour purity. The interaction term was used to identify genes with age-associated mRNA changes. FDR adjusted p-values and fold-changes were extracted for visualization. A q-value threshold of 0.1 was used for statistical significance. For genes affected by age-associated gains, the same procedure was applied using copy number gains. mRNA modeling results for age-associated CNAs and SNVs are in **Supplementary Tables 5-6**.

Survival analysis

Performed for TCGA data. Genes found to have significant (FDR threshold of 10%) age-associated CNAs and SNVs were also analyzed using Cox proportional hazards modelling after checking proportional hazards assumption. Cox proportional hazard regression models incorporating age, CNA/SNV status, and an age-CNA/SNV group interaction were fit for overall survival after checking the proportional hazards assumption. Age was treated as a continuous variable for modeling, but median dichotomized into 'low age' and 'high age' groups for visualization. FDR-adjusted interaction p-values and $\log_2$ hazard ratios were extracted for visualization. A q-value threshold of 0.1 was used to identify genes with sex-influenced survival. Survival modeling results for age-associated CNAs and SNVs are in **Supplementary Tables 5 and 6**.

Statistical Analysis & Data Visualization Software

All statistical analyses and data visualization were performed in the R statistical environment (v3.2.1) using the BPG[91] (v5.9.8) and Survival (v2.44-1.1) packages, and with Inkscape (v0.92.3).

References

1. Cook, M. B. *et al.* Sex disparities in cancer incidence by period and age. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **18**, 1174–1182 (2009).
2. Edgren, G., Liang, L., Adami, H.-O. & Chang, E. T. Enigmatic sex disparities in cancer incidence. *Eur. J. Epidemiol.* **27**, 187–196 (2012).
3. Elsaleh, H. *et al.* Association of tumour site and sex with survival benefit from adjuvant chemotherapy in colorectal cancer. *Lancet Lond. Engl.* **355**, 1745–1750 (2000).
4. Cook, M. B., McGlynn, K. A., Devesa, S. S., Freedman, N. D. & Anderson, W. F. Sex Disparities in Cancer Mortality and Survival. *Cancer Epidemiol. Biomarkers Prev.* **20**, 1629–1637 (2011).
5. Scelo, G., Li, P., Chanudet, E. & Muller, D. C. Variability of Sex Disparities in Cancer Incidence over 30 Years: The Striking Case of Kidney Cancer. *Eur. Urol. Focus* **4**, 586–590 (2018).
6. Zheng, D. *et al.* Sexual dimorphism in the incidence of human cancers. *BMC Cancer* **19**, 684 (2019).
7. Nipp, R. *et al.* Disparities in cancer outcomes across age, sex, and race/ethnicity among patients with pancreatic cancer. *Cancer Med.* **7**, 525–535 (2018).
8. Zhang, W., Edwards, A., Flemington, E. K. & Zhang, K. Racial disparities in patient survival and tumor mutation burden, and the association between tumor mutation burden and cancer incidence rate. *Sci. Rep.* **7**, 13639 (2017).
9. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA. Cancer J. Clin.* **70**, 7–30 (2020).

10. Torre, L. A. *et al.* Cancer statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016: Converging incidence in males and females: Cancer Statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016. *CA. Cancer J. Clin.* **66**, 182–202 (2016).
11. DeSantis, C. E., Miller, K. D., Goding Sauer, A., Jemal, A. & Siegel, R. L. Cancer statistics for African Americans, 2019. *CA. Cancer J. Clin.* **69**, 211–233 (2019).
12. Cook, P. J., Doll, R. & Fellingham, S. A. A mathematical model for the age distribution of cancer in man. *Int. J. Cancer* **4**, 93–112 (1969).
13. de Magalhães, J. P. How ageing processes influence cancer. *Nat. Rev. Cancer* **13**, 357–365 (2013).
14. Harding, C., Pompei, F. & Wilson, R. Peak and decline in cancer incidence, mortality, and prevalence at old ages. *Cancer* **118**, 1371–1386 (2012).
15. White, M. C. *et al.* Age and Cancer Risk. *Am. J. Prev. Med.* **46**, S7–S15 (2014).
16. DePinho, R. A. The age of cancer. *Nature* **408**, 248 (2000).
17. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **8**, 1–12 (1954).
18. Ershler, W. B. Cancer: a disease of the elderly. *J. Support. Oncol.* **1**, 5–10 (2003).
19. Aunan, J. R., Cho, W. C. & Søreide, K. The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging Dis.* **8**, 628 (2017).
20. Steliarova-Foucher, E. *et al.* International incidence of childhood cancer, 2001–10: a population-based registry study. *Lancet Oncol.* **18**, 719–731 (2017).

21. the St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nat. Genet.* **45**, 602–612 (2013).
22. the St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project. The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat. Genet.* **46**, 444–450 (2014).
23. Trama, A. *et al.* Survival of European adolescents and young adults diagnosed with cancer in 2000–07: population-based data from EUROCARE-5. *Lancet Oncol.* **17**, 896–906 (2016).
24. Anders, C. K. *et al.* Young Age at Diagnosis Correlates With Worse Prognosis and Defines a Subset of Breast Cancers With Shared Patterns of Gene Expression. *J. Clin. Oncol.* **26**, 3324–3330 (2008).
25. Tingstedt, B., Weitkämper, C. & Andersson, R. Early onset pancreatic cancer: a controlled trial. *Ann. Gastroenterol.* **24**, 206–212 (2011).
26. Piciucchi, M. *et al.* Early onset pancreatic cancer: Risk factors, presentation and outcome. *Pancreatology* **15**, 151–155 (2015).
27. Katz, M. *et al.* The Effect of Race/Ethnicity on the Age of Colon Cancer Diagnosis. *J. Health Disparities Res. Pract.* **6**, 62–69 (2013).
28. Tricoli, J. V. *et al.* Biologic and clinical characteristics of adolescent and young adult cancers: Acute lymphoblastic leukemia, colorectal cancer, breast cancer, melanoma, and sarcoma: Biology of AYA Cancers. *Cancer* **122**, 1017–1028 (2016).
29. Willauer, A. N. *et al.* Clinical and molecular characterization of early-onset colorectal cancer. *Cancer* **125**, 2002–2010 (2019).

30. Lieu, C. H. *et al.* Comprehensive Genomic Landscapes in Early and Later Onset Colorectal Cancer. *Clin. Cancer Res.* **25**, 5852–5858 (2019).
31. Parsons, D. W. *et al.* An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science* **321**, 1807–1812 (2008).
32. Williams, G. R. *et al.* Comorbidity in older adults with cancer. *J. Geriatr. Oncol.* **7**, 249–257 (2016).
33. Jørgensen, T. L., Hallas, J., Friis, S. & Herrstedt, J. Comorbidity in elderly cancer patients in relation to overall and cancer-specific mortality. *Br. J. Cancer* **106**, 1353–1360 (2012).
34. Balducci, L., Colloca, G., Cesari, M. & Gambassi, G. Assessment and treatment of elderly patients with cancer. *Surg. Oncol.* **19**, 117–123 (2010).
35. Given, B. & Given, C. W. Older adults and cancer treatment. *Cancer* **113**, 3505–3511 (2008).
36. Chen, R. C., Royce, T. J., Extermann, M. & Reeve, B. B. Impact of Age and Comorbidity on Treatment and Outcomes in Elderly Cancer Patients. *Semin. Radiat. Oncol.* **22**, 265–271 (2012).
37. Andaya, A. A. *et al.* Race and Colon Cancer Survival in an Equal-Access Health Care System. *Cancer Epidemiol. Biomarkers Prev.* **22**, 1030–1036 (2013).
38. Milholland, B., Auton, A., Suh, Y. & Vijg, J. Age-related somatic mutations in the cancer genome. *Oncotarget* **6**, 24627–24635 (2015).
39. Wahl, D. R. *et al.* Pan-Cancer Analysis of Genomic Sequencing Among the Elderly. *Int. J. Radiat. Oncol.* **98**, 726–732 (2017).

40. Li, C. H., Haider, S., Shiah, Y.-J., Thai, K. & Boutros, P. C. Sex Differences in Cancer Driver Genes and Biomarkers. *Cancer Res.* **78**, 5527 (2018).
41. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, (2017).
42. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
43. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
44. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
45. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
46. Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
47. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
48. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).
49. Vollan, H. K. M. *et al.* A tumor DNA complex aberration index is an independent predictor of survival in breast and ovarian cancer. *Mol. Oncol.* **9**, 115–127 (2015).
50. Lalonde, E. *et al.* Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol.* **15**, 1521–1532 (2014).

51. Hieronymus, H. *et al.* Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11139–11144 (2014).
52. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124–128 (2015).
53. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
54. PCAWG Tumour Subtypes and Clinical Translation *et al.* Sex differences in oncogenic mutational processes. *Nat. Commun.* **11**, 4330 (2020).
55. PCAWG Evolution & Heterogeneity Working Group *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
56. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* **173**, 1003-1013.e15 (2018).
57. Landau, D. A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152**, 714–726 (2013).
58. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
59. Asomaning, K. *et al.* Second hand smoke, age of exposure and lung cancer risk. *Lung Cancer* **61**, 13–20 (2008).
60. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
61. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).

62. PCAWG Mutational Signatures Working Group *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
63. Alexandrov, L. *et al.* The Repertoire of Mutational Signatures in Human Cancer. (2018) doi:10.1101/322859.
64. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
65. PCAWG Drivers and Functional Interpretation Working Group *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
66. Stewart, J. B. *et al.* Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers. *PLOS Genet.* **11**, e1005333 (2015).
67. Hopkins, J. F. *et al.* Mitochondrial mutations drive prostate cancer aggression. *Nat. Commun.* **8**, 656 (2017).
68. Ye, C. *et al.* Mutations in the mitochondrial DNA D-loop region and breast cancer risk. *Breast Cancer Res. Treat.* **119**, 431–436 (2010).
69. Nugent, S. *et al.* Brain glucose and acetoacetate metabolism: a comparison of young and older adults. *Neurobiol. Aging* **35**, 1386–1395 (2014).
70. van den Bent, M. J., Smits, M., Kros, J. M. & Chang, S. M. Diffuse Infiltrating Oligodendroglioma and Astrocytoma. *J. Clin. Oncol.* **35**, 2394–2401 (2017).
71. Li, C. H. *et al. Sex Differences in Oncogenic Mutational Processes*. http://biorxiv.org/lookup/doi/10.1101/528968 (2019) doi:10.1101/528968.
72. Chatsirisupachai, K., Lesluyes, T., Paraoan, L., Van Loo, P. & de Magalhães, J. P. *An Integrative Analysis of the Age-Associated Genomic, Transcriptomic and Epigenetic*

Landscape across Cancers. http://biorxiv.org/lookup/doi/10.1101/2020.08.25.266403 (2020) doi:10.1101/2020.08.25.266403.

73. Maynard, S., Fang, E. F., Scheibye-Knudsen, M., Croteau, D. L. & Bohr, V. A. DNA Damage, DNA Repair, Aging, and Neurodegeneration. *Cold Spring Harb. Perspect. Med.* **5**, a025130 (2015).
74. Hoeijmakers, J. H. J. DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* **361**, 1475–1485 (2009).
75. Fernández, L. C., Torres, M. & Real, F. X. Somatic mosaicism: on the road to cancer. *Nat. Rev. Cancer* **16**, 43–55 (2016).
76. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
77. Özdemir, B. C. & Dotto, G.-P. Racial Differences in Cancer Susceptibility and Survival: More Than the Color of the Skin? *Trends Cancer* **3**, 181–197 (2017).
78. Cramer, D. W. & Finn, O. J. Epidemiologic perspective on immune-surveillance in cancer. *Curr. Opin. Immunol.* **23**, 265–271 (2011).
79. Vick, A. D. & Burris, H. H. Epigenetics and Health Disparities. *Curr. Epidemiol. Rep.* **4**, 31–37 (2017).
80. Huang, K. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370.e14 (2018).
81. Research, W. C. R. F. I. for C. Diet, nutrition, physical activity and cancer: a global perspective. *Contin. Update Proj. Expert Rep.* (2018).

82. Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A. & Wargo, J. A. The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. *Cancer Cell* **33**, 570–580 (2018).
83. Bostwick, D. G. *et al.* Human prostate cancer risk factors. *Cancer* **101**, 2371–2490 (2004).
84. Irigaray, P. & Belpomme, D. Basic properties and molecular mechanisms of exogenous chemical carcinogens. *Carcinogenesis* **31**, 135–148 (2010).
85. Talarico, L., Chen, G. & Pazdur, R. Enrollment of Elderly Patients in Clinical Trials for Cancer Drug Registration: A 7-Year Experience by the US Food and Drug Administration. *J. Clin. Oncol.* **22**, 4626–4631 (2004).
86. Ruiter, R., Burggraaf, J. & Rissmann, R. Under-representation of elderly in clinical trials: An analysis of the initial approval documents in the Food and Drug Administration database. *Br. J. Clin. Pharmacol.* **85**, 838–844 (2019).
87. Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell* **37**, 639-654.e6 (2020).
88. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
89. The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
90. Yeo, I.-K. & Johnson, R. A. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* **87**, 954–959 (2000).
91. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics* **20**, (2019).

Acknowledgments

The authors thank all the members of the Boutros lab for insightful discussion. We also thank João Pedro de Magalhães, Kasit Chatsirisupachai, Peter Van Loo, and the de Magalhães lab for collaborative discussion. This work was supported by the Discovery Frontiers: Advancing Big Data Science in Genomics Research program, which is jointly funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), Genome Canada and the Canada Foundation for Innovation (CFI). P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. This work was supported by an NSERC Discovery grant and by Canadian Institutes of Health Research, grant #SVB-145586, to PCB. The results described here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This work was supported by the NIH/NCI under award number P30CA016042 and by an operating grant from the National Cancer Institute Early Detection Research Network (U01CA214194).

Author Contributions

CHL and PCB initiated the project. CHL, and SH analyzed data. PCB supervised research. CHL and PCB wrote the first draft of the manuscript, which all authors edited and approved.

Figures & Figure Legends

Figure 1

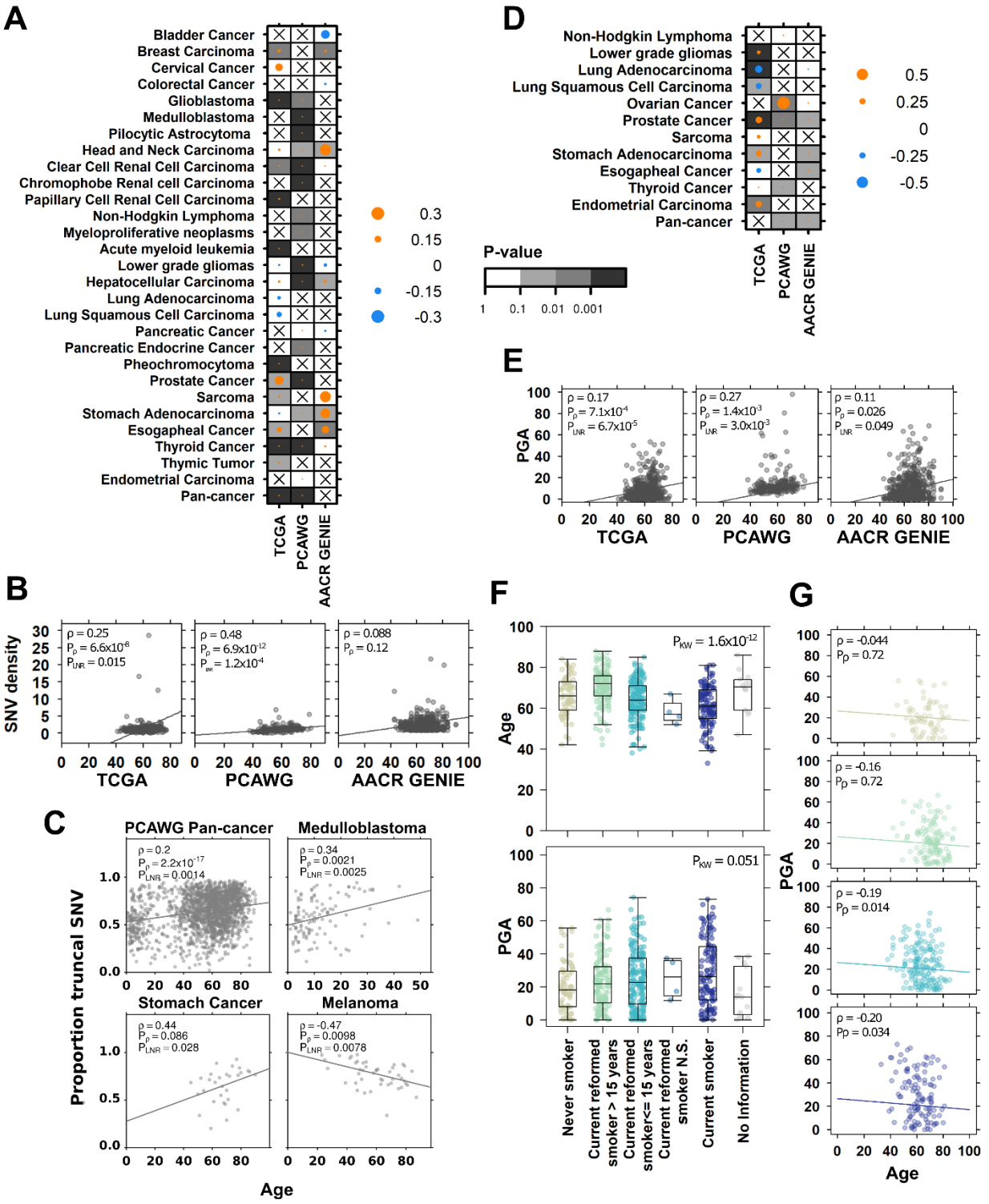

Figure 1 | Mutation density and timing are associated with age at diagnosis. Summary of associations between age and **(A)** SNV density and **(D)** PGA in TCGA, PCAWG, and AACR GENIE tumours. The dot size and colour show the Spearman correlation, and background shading indicate adjusted multivariate p-value. Only tumour-types with at least univariately significant associations are shown. Associations between **(B)** SNV density and **(E)** PGA with age in prostate cancer across the three datasets. Univariate Spearman correlation, adjusted correlation p-value and adjusted multivariate p-values shown. **(C)** Correlations between age and proportion of SNVs occurring in the truncal clone in four PCAWG tumour contexts. **(F)** In TCGA lung adenocarcinoma, age and PGA are associated with smoking history, and **(G)** the negative association between PGA and age remains significant in current smokers and recently reformed smokers (<=15 years). Dot colours show smoking history groups as indicated in **(F)**.

Figure 2

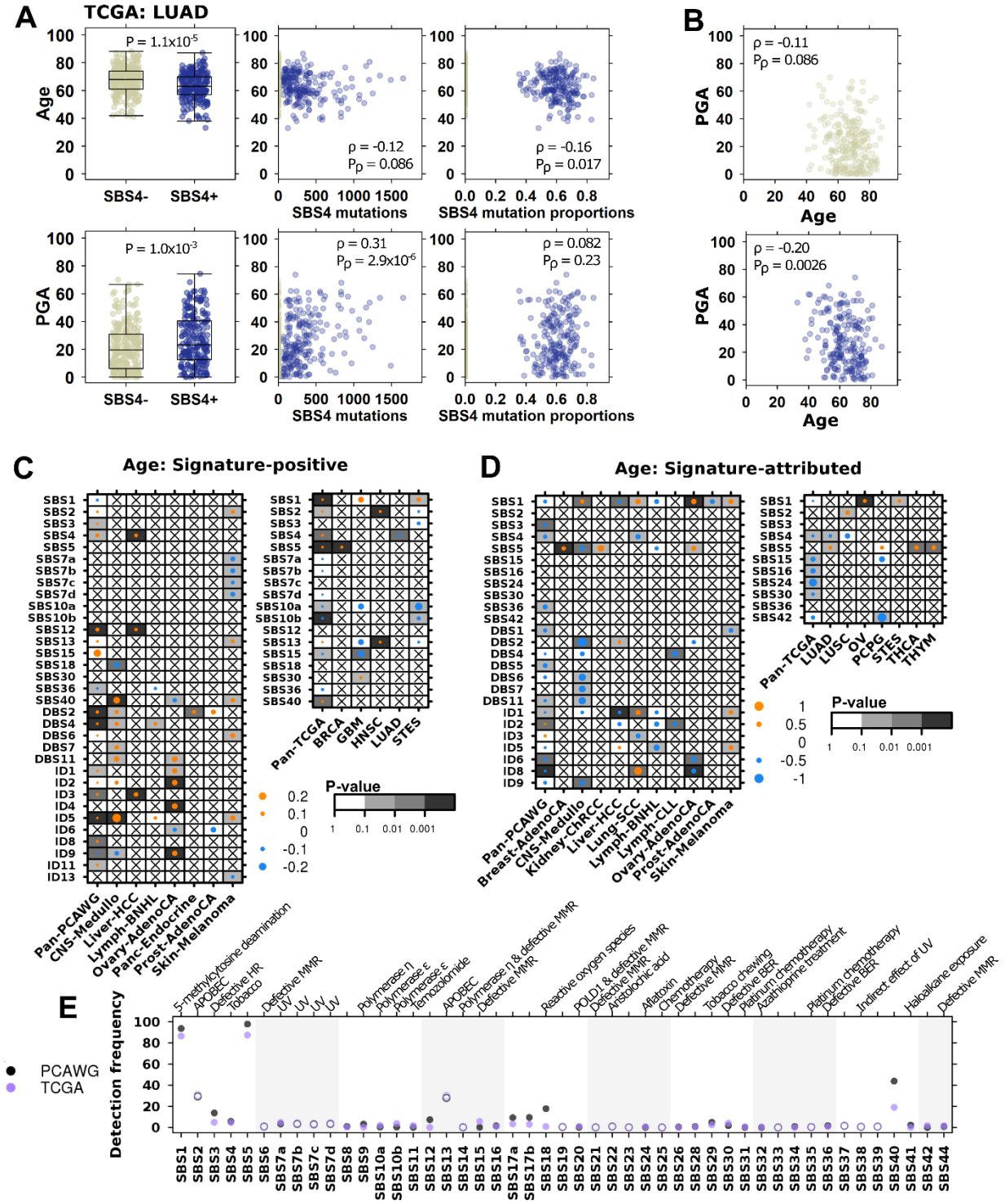

Figure 2 | Age-associated mutational signatures suggest differences in underlying mutational processes.

(A) In TCGA lung adenocarcinoma, age (top plots) and PGA (bottom plots) are associated with the tobacco-associated signature SBS4 (left). The absolute number of SBS4-attributed mutations is also associated with age and PGA (middle). The relative proportion of SBS4-attibuted mutations is negatively associated with age and has no significant relationship with PGA (right). **(B)** The negative association between PGA and age remains significant in both SBS4+ and SBS4- groups. Yellow and blue dots indicate SBS4- and SBS4+ tumours, respectively. **(C)** Summary of associations between age and the proportion of signature-positive tumours, where dot size shows the marginal log odds from logistic regression and background shading show adjusted multivariate p-values. PCAWG data is on left and TCGA on right. **(D)** Similarly, the summary of associations between age and relative signature activity, with dot size showing Spearman correlations and background indicating adjusted linear regression p-values. **(E)** Comparison of PCAWG and TCGA signature detection frequency. Filled in and open circles indicate comparisons where the differences are statistically significant (proportion test $q < 0.05$) and not, respectively. Proposed SBS signature aetiologies are as indicated. Proposed DBS and ID aetiologies are: DBS1: UV, DBS2: tobacco, DBS5: platinum chemotherapy, DBS7: defective MMR, ID1: slippage during DNA replication, ID2: slippage during DNA replication, ID3: tobacco, ID6: defective HR, ID8: NHEJ, ID13: UV

Figure 3

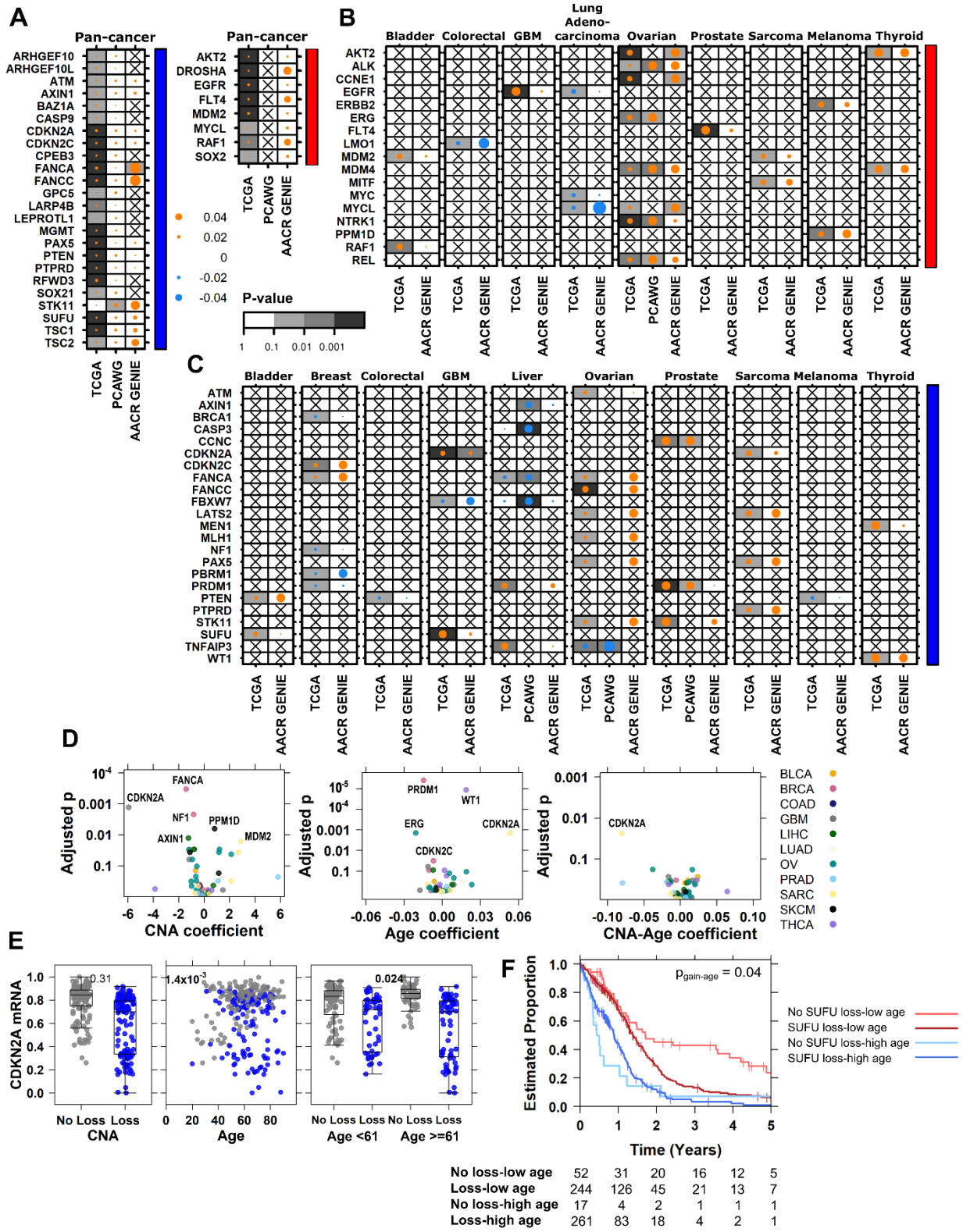
A
Pan-cancer
ARHGEF10
ARHGEF10L
ATM
AXIN1
BAZ1A
CASP9
CDKN2A
CDKN2C
CPEB3
FANCA
FANCC
GPC5
LARP4B
LEPROTL1
MGMT
PAX5
PTEN
PTPRD
RFWD3
SOX21
STK11
SUFU
TSC1
TSC2
TCGA
PCAWG
AACR GENIE
Pan-cancer
AKT2
DROSHA
EGFR
FLT4
MDM2
MYCL
RAF1
SOX2
0.04
0.02
0
-0.02
-0.04
P-value
1
0.1
0.01
0.001
B
Bladder
Colorectal
GBM
Lung Adeno-carcinoma
Ovarian
Prostate
Sarcoma
Melanoma
Thyroid
AKT2
ALK
CCNE1
EGFR
ERBB2
ERG
FLT4
LMO1
MDM2
MDM4
MITF
MYC
MYCL
NTRK1
PPM1D
RAF1
REL
C
Bladder
Breast
Colorectal
GBM
Liver
Ovarian
Prostate
Sarcoma
Melanoma
Thyroid
ATM
AXIN1
BRCA1
CASP3
CCNC
CDKN2A
CDKN2C
FANCA
FANCC
FBXW7
LATS2
MEN1
MLH1
NF1
PAX5
PBRM1
PRDM1
PTEN
PTPRD
STK11
SUFU
TNFAIP3
WT1
D
Adjusted p
FANCA
CDKN2A
NF1
PPM1D
AXIN1
MDM2
CNA coefficient
PRDM1
WT1
ERG
CDKN2A
CDKN2C
Age coefficient
CDKN2A
CNA-Age coefficient
BLCA
BRCA
COAD
GBM
LIHC
LUAD
OV
PRAD
SARC
SKCM
THCA
E
CDKN2A mRNA
0.31
1.4x10-3
0.024
No Loss Loss
CNA
Age
No Loss Loss No Loss Loss
Age <61
Age >=61
F
Estimated Proportion
pgain-age = 0.04
No SUFU loss-low age
SUFU loss-low age
No SUFU loss-high age
SUFU loss-high age
Time (Years)
No loss-low age 52 31 20 16 12 5
Loss-low age 244 126 45 21 13 7
No loss-high age 17 4 2 1 1 1
Loss-high age 261 83 18 4 2 1

Figure 3 | Age-associations in copy number drivers are associated with functional changes in mRNA and survival.

Summary of all detected age-associated driver CNAs for **(A)** pan-cancer and **(B-C)** tumour-type specific analyses across three datasets. Dot size shows the magnitude of the association as the difference in proportion and the background shading shows adjusted multivariate p-values. Right covariate indicates copy number gain drivers in red and loss drivers in blue. **(D)** Age-associated CNAs lead to differential mRNA abundance with respect to the CNA itself (left), age (middle), as well as age-specific effects of the CNA (right). **(E)** *CDKN2A* mRNA abundance changes between copy number loss (blue) or no loss (black) and broken down by age for low *vs*. high age. Tukey boxplots are shown with the box indicating quartiles and the whiskers drawn at the lowest and highest points within 1.5 interquartile range of the lower and upper quartiles, respectively. Adjusted p-values are indicated. **(F)** *SUFU* loss interacts with age to further stratify patient prognosis. The adjusted p-value for the copy number loss-age interaction term is shown.

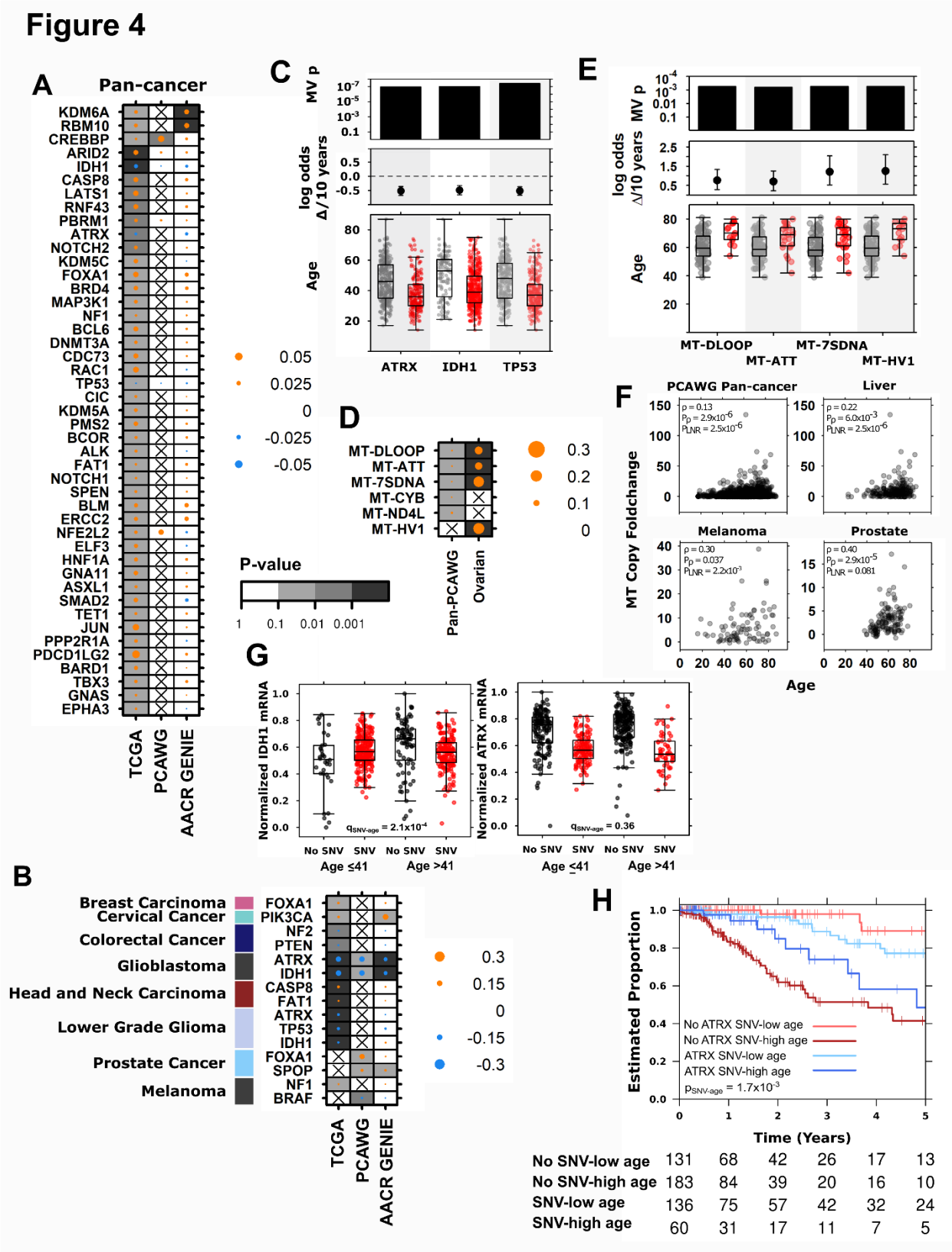

Figure 4 | Age-associations in nuclear and mitochondrial SNVs reveal ATRX as an

age-associated prognostic biomarker in lower grade glioma.
(A) Summary of age- associated nuclear driver SNVs for **(A)** pan-cancer, **(B)** tumour-type specific analyses across three datasets. Dot size shows the magnitude of the association as the difference in proportion and the background shading shows adjusted multivariate p-values. Left covariate in **(B)** indicates relevant tumour-type. **(C)** TCGA lower grade glioma age-associations in driver mutation frequency with adjusted multivariate p-values, marginal log odds changes for 10-year age increment, and age of tumours compared between those with (red) and without (grey) the mutation. **(D)** Summary of age-associated mitochondrial SNVs in PCAWG data with specific examples from ovarian cancer shown in **(E)**. **(F)** Mitochondrial copy number foldchange is also associated with age in four tumour contexts. **(G)** In TCGA lower grade glioma: mRNA abundance changes for *IDH1* and *ATRX* when the gene is mutated (red) or not (black) compared by median-dichotomized age. Adjusted SNV-age interaction p-values are shown. **(H)** *ATRX* mutation interacts with age to stratify patient prognosis into four groups. Log-odds p-value is shown. Tukey boxplots are shown with the box indicating quartiles and the whiskers drawn at the lowest and highest points within 1.5 interquartile range of the lower and upper quartiles, respectively.