Age Influences on the Molecular Presentation of Tumours

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

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

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-biases in the somatic mutational landscape of 12,774 tumours across 33 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. A subset of known cancer driver genes were mutated in age-biased patterns, and these alter the transcriptome and predict for clinical outcomes. These effects were most striking in lower grade glioma where *ATRX* mutation is a strongly age-dependent prognostic biomarker. Though cancer genome sequencing data is not well-balanced in epidemiologic factors, these data suggest that age shapes the somatic mutational landscape of cancer, with clear 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 associated with higher mortality and lower survival[16,17]. The links between older age and increased cancer burden 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 prostate[23], breast[24], and colorectal[25] cancers 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[26–28], suggesting differences in the underlying oncogenic processes driving cancer at different ages.

The mechanisms of how age shapes the clinical behaviour of cancers has 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[29,30]. 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[31–33]. Nevertheless differences remain even after accounting for these factors[34]. Previous work associating somatic molecular changes with age suggest differences in overall tumour mutation burden[35], transcriptional profiles[36], and some mutational differences[26–28]. 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 perform 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) and 2,562 tumours of 30 tumour-types from the International Cancer Genome Consortium/The Cancer Genome Atlas Pan-cancer Analysis of Whole Genomes (PCAWG) projects. We quantified age-biases in measures of mutation density, subclonal architecture, mutation timing, mutational signatures and driver mutations in almost all tumour-types. We adjusted for potential confounding factors such as sex and ancestry. Many of these genomic age-biases 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 Biases in Mutation Density and Timing

We investigated TCGA and PCAWG datasets independently and performed pan-cancer analyses spanning all TCGA tumours (pan-TCGA), and all PCAWG tumours (pan-PCAWG); these were supplemented with tumour-type-specific analyses. We used the recorded age at diagnosis for both TCGA and PCAWG[37] (**Table 1**). Our modeling accounted for a range of confounding variables for each cancer type including sex and genetic ancestry. We adapted a statistical approach previously applied to quantify sex-biases in cancer genomics[38]: we first used univariate methods to identify putative age-biases, then further modeled these putative hits with multivariate regression to evaluate age effects after adjusting for confounding factors. 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 began by assessing whether measures of mutation density were associated with age. The accumulation of mutations with age is a well-known phenomenon in both cancer and non-cancer cells[39–46]. We examined both genome instability and SNV density to investigate trends across age and test the robustness of our statistical framework in detecting age-associated genomic events. Genome instability is a measure of copy number aberration (CNA) burden and approximated by percent of the genome altered by CNAs (PGA), a surrogate variable associated with poor outcome in several tumour-types[47–49]. We identified univariate age-biases in PGA 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 discovered significant associations between age and PGA in both pan-TCGA ($\rho = 0.14$, adjusted LNR $p = 1.1 \times 10^{-7}$) and pan-PCAWG ($\rho = 0.19$, adjusted LNR $p = 0.023$) data. Positive correlations were also identified in three TCGA and three PCAWG tumour-types, with prostate cancer showing a statistically significant correlation in both datasets. (**Figure 1A, 1B, Supplementary Table 2**). Other tumour-type specific associations were statistically significant in only one dataset (**Figure 1A**, **Supplementary Figure 1**). For example, we detected similar correlations between age and PGA in TCGA lower grade glioma (LGG) and PCAWG (CNS-Oligo), but the association was significant in only TCGA data (**Figure 1A, 1B**). This is likely due in part to decreased statistical power in the PCAWG dataset because of smaller sample sizes. Surprisingly, in TCGA both adenocarcinomas and squamous cell carcinomas of the lung showed the inverse trend, with tumours arising in older patients harbouring fewer CNAs (LUAD: $\rho = -0.18$, adjusted

LNR p = 6.0×10^{-4}, LUSC: ρ = -0.10, adjusted LNR p = 0.039, **Figure 1A**). We observed similar negative correlations in the corresponding PCAWG lung data, though these associations were not statistically significant (Lung-AdenoCA: ρ = -0.13, Lung-SCC: ρ = -0.065, **Figure 1A**).

Analogous to PGA, somatic single nucleotide variation (SNV) density measures the burden of somatic SNVs. SNV density frequently increased with age, as expected[39,40]. In addition to pan-cancer age-biases (pan-TCGA: ρ = 0.33, adjusted LNR p = 2.0×10^{-49}, pan-PCAWG: ρ = 0.41, adjusted LNR p = 3.1×10^{-28}), tumour-type-specific positive correlations occurred in 15/23 TCGA and 14/30 PCAWG tumour-types, including in prostate and gastric cancers (**Figure 1C**, **1D Supplementary Figure 1, Supplementary Table 2**). Again, there was an inverse relationship in lung tumours, with more SNVs occurring in the squamous cell tumours of younger patients (LUSC: ρ = -0.15, adjusted LNR p = 0.064, **Figure 1D**). While not statistically significant, we observed similar negative associations in PCAWG lung tumours (Lung-SCC: ρ = -0.14). The negative association between age and both PGA and SNV density in lung cancers has been attributed to smoking exposure leading to hypermutation in younger lung cancer patients[50], suggesting differences in disease aetiology between patients of different ages.

Another source of of hypermutation is microsatellite instability (MSI), which is frequently detected in colorectal and gastric cancers[51,52]. Since MSI-positive status is often associated with increased SNV density and age (**Supplementary Figure 2**), we investigated whether age-biases in MSI might explain the associations between age and SNV density in this data. We focused on four tumour-types with high frequency of MSI-positive tumours: stomach & esophageal, colorectal, pancreatic and endometrial cancers[53,54]. There was a significant association between age and MSI status in gastric cancers, where tumours arising in older individuals were more likely to have high levels of MSI (MSI-H; ANOVA q = 6.4×10^{-4}; **Supplementary Figure 2**). While there were no statistically significant associations between age and MSI status in colorectal, pancreatic, or endometrial cancers, we nevertheless assessed the relationship between age and SNV density while accounting for MSI status in all four tumour-types. The associations between age and SNV density remained significant even after adjusting for MSI status in stomach & esophageal, colorectal, and pancreatic tumours (**Supplementary Figure 2**).

After identifying age-biases in mutation density, we next asked whether there were differences in the timing of when these mutations occurred during tumour evolution. We leveraged data describing the evolutionary history of PCAWG tumours[55] and first investigated polyclonality, or the number of cancer cell populations detected in each tumour as assessed by multiple methods in this dataset. 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]. We also investigated mutation timing in polyclonal tumours by comparing how frequently SVNs, indels and structural variants (SVs) occurred

as clonal mutations in the trunk or as subclonal ones in branches. While there were intriguing univariate associations between age and polyclonality in non-Hodgkin lymphoma and prostate cancer, they were not significant after multivariate adjustment (**Supplementary Figure 2**).

Focusing on polyclonal tumours, we compared how frequently mutations occurred in the trunk subclone *vs*. in branch subclones. Differences in the proportion of truncal mutations suggest difference in mutation timing over the evolution of a tumour. 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 1E**) 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 1E**). A positive correlation results suggest that in these tumour-types, tumours arising in older individuals accumulate a greater fraction of SNVs earlier in tumour evolution. In contrast, an inverted trend occurred in melanoma, where tumours of younger patients tended to accumulate more subclonal than clonal SNVs (ρ = -0.47, adjusted LNR p = 7.8×10^{-3}).

Age Biases in Mutational Processes

Differences in mutation density and timing suggest that different oncogenic processes might be preferentially active depending on the age of a patient. These processes can result in distinctive mutational patterns, which can be deconvolved and quantified[59]. We analysed age-biases in three types of mutational signatures generated by the PCAWG project: 49 single base substitution (SBS), 11 doublet base substitution (DBS) and 17 small insertion and deletion (ID) signatures[60]. We also investigated SBS signatures for TCGA tumours. For each signature, we examined both the proportion of signature-positive tumours as well as relative mutation activity, or the proportion of mutations attributed to each signature.

Across all 2,562 PCAWG tumours, we identified twelve mutational signatures with age-biased detection frequency (**Figure 2A, left**) and ten with age-biased mutation activity (**Figure 2B, 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 (ρ = -0.20, adjusted LNR p = 3.2×10^{-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 x 10^{-3}; ρ = -0.099, adjusted LNR p = 3.7 x 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; ρ = -0.059, adjusted LNR p = 0.048). We also identified positive associations between higher age and the tobacco-related signatures SBS4, 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 2A**), attributed to APOBEC cytidine deaminase activity[61]. Melanomas arising in younger patients were more likely to be positive for signatures related to UV damage (SBS 7a, 7b, 7d, **Figure 2A, Supplementary Table 2**). The proportion of mutations attributed to UV damage was also higher in younger patients (DBS1, ρ = -0.29, adjusted LNR p = 0.019, **Figure 2B**), while the proportion of mutations attributed to slippage during DNA replication was higher in older patients (ID1, ρ = 0.27, adjusted LNR p = 0.019, **Figure 2B**). 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 2A**). We also identified five signatures with higher relative activity in younger patients (**Figure 2B**). In cancer-specific analysis, we identified age-biased SBS signatures across eleven tumour-types, including negative associations between the tobacco-associated signature SBS4 and age in lung adenocarcinoma. SBS4 was more frequently detected in younger patients (LUAD: marginal log odds change = -0.041, 95%CI = -0.062 - -0.021, adjusted LGR p = 4.2 x 10^{-3}, **Figure 2A**) and also had higher relative activity in younger lung squamous cell cancer patients (ρ = -0.17, adjusted LNR p = 0.015, **Figure 2B**). SBS4 activity was similarly negatively associated with age in PCAWG lung squamous cell cancers (Lung-SCC: ρ = -0.35, adjusted LNR p = 0.099, **Figure 2B**). 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[50].

There was moderate agreement between TCGA and PCAWG findings: some signature like SBS2 and SBS4 were age-biased in the same or closely related tumour subtypes. Other signatures, such as SBS1 and SBS5 were age-biased in detection and relative activity across a range of tumour-types. Still others were age-biased exclusively in either TCGA or PCAWG data. 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 2C**). 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 3**). We also looked specifically at identified age-biases and found high agreement in data generated by the same sequencing strategy (**Supplementary Figure 3**). The differences in signature detection, sequencing strategy, multivariate models, sample size, and geographic variation distinguishing PCAWG and TCGA datasets motivated our continued analysis of each dataset separately.

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 chromosome segment and gene-level events. We applied our statistical framework to identify putative age-biased copy number gains and losses using univariate logistic regression (ULR). 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 and TCGA datasets separately to characterize pan-cancer and tumour-type-specific biases. Pan-PCAWG, we identified 1,158 genes in age-associated CNAs (**Figure 3A**, **Supplementary Figure 4**). All significant age-biased losses occurred more frequently in older patients. In pan-TCGA data, we identified a greater number of age-biased events with 8,583 genes in age-associated losses and 15,497 genes in age-associated gains (**Figure 3A, 3B**). These global pan-cancer age-associations were reflected in 17 individual TCGA and four PCAWG tumour-types (**Figure 3A**). PCAWG and TCGA analyses were well-correlated, for example as seen in ovarian cancer (**Supplementary Figure 4**). We further focused on a set of 133 genes with driver CNAs[62]. Across all cancer types, we identified 64 drivers with positive CNA-age associations (41 gains, 23 losses, **Figure 3C**). In tumour-type specific analysis, we

found age-associated driver CNAs in 16 TCGA and 5 PCAWG tumour-types (**Supplementary Tables 3-4**).

We next asked whether statistically significant age-biased CNAs perturb the transcriptome 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. In glioblastoma our CNA analysis identified 3,829 genes in age-associated gains and 3,754 genes in age-associated losses (**Figure 3D, Supplementary Tables 3-4**). For example, *DBNDD2* was more frequently gained (marginal log odds change = 0.025, 95%CI = 0.013-0.037, adjusted MLR $p = 6.6 \times 10^{-5}$) and *RASSF4* was more frequently lost (marginal log odds change = 0.059, 95%CI = 0.043-0.076, adjusted MLR $p = 8.5 \times 10^{-11}$) in tumours of older individuals (**Figure 3E**).

Of these age-biased CNAs, we identified 379 genes with significant copy number-mRNA associations and 27 with significant CNA-age interactions (**Figure 3F**). For instance, gain of the gene *DBNDD2* is itself associated with increased mRNA abundance (adjusted CNA $p = 1.2 \times 10^{-3}$), but there is also a strong age-dependent effect: *DBNDD2* gain is associated with increased mRNA abundance in tumours arising in older individuals, but decreased mRNA abundance in tumours of younger individuals, (adjusted CNA-age $p = 3.5 \times 10^{-3}$, **Figure 3G**). Other examples of this significant interaction include loss of *SMPD1* which affects mRNA abundance more in tumours arising in younger patients (**Supplementary Figure 4**, **Supplementary Table 5**). Thus, these age-biased CNAs not only preferentially occur in tumours derived of individuals of a certain age group, they are also associated with changes in mRNA that are often biased by age as well. These data highlight the complex interplay between CNAs, age and 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 overall survival as the end point. Similar to our mRNA models, we used predictors including copy number status, age and their interaction (**Figure 3H**). 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 $p = 1.4 \times 10^{-13}$). We found that loss of a segment on chromosome 10q containing 31 genes including *RASSF4* and *RSU1P2* was also prognostic, with loss of this region associated with poorer survival (HR = 1.9, 95%CI = 1.4 - 2.7, adjusted Wald $p = 2.6 \times 10^{-3}$). Integrating age and this 10q segment loss reveals three groups with distinct survival trajectories: older individuals have the worst outcomes regardless of copy number loss status, but younger individuals with the loss have poorer outcome than those without it (**Figure 3I**). We performed survival modeling for 5,251

genes on affected by age-biased CNAs in glioblastoma and found 1,821 genes showed associations between copy number change and prognosis and 142 genes had significant CNA-age interactions.

We repeated these mRNA and survival analyses for all TCGA tumour-types with age-biased CNAs. Genes in biased CNAs were associated with altered mRNA abundance in 12 tumour-types and interacted with age in age-dependent mRNA change in seven tumour-types. We observed a range of synergistic and antagonistic interactions for both gains and losses. In most cases, an age-biased copy number altered gene was associated with a greater mRNA abundance change in tumours of younger patients. In lower grade glioma, we observed synergy between CNA and age where the change in mRNA abundance was greatest in tumours of older patients. Six tumour-types also showed that biased 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-biases. In PCAWG analysis, we used a predefined set of driver mutations[63]. In TCGA analysis, we used a recurrent threshold to filter out genes mutated in less than 1% of tumours. We included SNV density in our multivariate models in addition to other confounding factors as previously described. We identified 15 age-biased genes across six PCAWG tumour contexts (**Figure 4A**), including a pan-cancer association with *CREBBP* (marginal log odds change = 0.027, 95%CI = 0.0089 – 0.047, adjusted LGR p = 8.7×10^{-3}). *CREBBP* was also associated with age in pan-cancer TCGA analysis (marginal log odds change = 0.032, 95%CI = 0.024 – 0.040, adjusted LGR p = 0.055). Pan-TCGA, we identified 401 genes that were mutated more frequently in older patients and four that were mutated more frequently in younger patients (**Supplementary Table 6**).

There were also tumour-type specific age-biases in SNV frequency, including age-associations of oncogenic *BRAF* mutations in PCAWG melanoma (marginal log odds change = -0.043, 95%CI = -0.072 - -0.017, adjusted LGR p = 2.4×10^{-3}), and *TERT* promoter mutations in PCAWG thyroid cancer (marginal log odds change = 0.10, 95%CI = 0.044 - 0.18, LGR p = 0.016). Age-biases in PCAWG medulloblastoma highlighted differences between paediatric and adult cases (**Figure 4A**), and tumours arising in older PCAWG prostate cancer patients were more likely to harbour *FOXA1* (marginal log odds change = 0.11, 95%CI = 0.041 - 0.18, adjusted LGR p = 0.013) and *SPOP* (marginal log odds change = 0.099, 95%CI = 0.032 - 0.18, adjusted LGR p = 0.060) mutations. We also confirmed known associations between lower age and mutations in tumour suppressors *IDH1* and *ATRX*, which were mutated in the same patients in PCAWG glioblastoma (marginal log odds change = -0.15, 95%CI = -0.31 - -0.052, adjusted LGR p = 0.017).

Similarly in TCGA data, we found higher frequency of *IDH1* and *ATRX* mutations in glioblastomas (**Figure 4B**) and lower grade gliomas of younger patients (**Figure 4C**). Other age-biased mutations occurred in pan-TCGA analysis for breast cancer, head & neck cancer, and stomach & esophageal cancer (**Supplementary Table 6**).

As with the age-biased CNAs, we evaluated the impact of SNVs on mRNA abundance and survival in TCGA data. We identified significant associations between age-biased SNVs and mRNA abundance for *ATRX* and *IDH1* in lower grade glioma (**Supplementary Table 6**). Mutations in *ATRX* and *IDH1* were associated with lowered mRNA abundance in both genes. There was also a significant interaction effect between age and *IDH1* mutation (adjusted $p = 9.7 \times 10^{-4}$, **Figure 4D**) indicating an age-dependent effect on mRNA abundance: mutated *IDH1* was associated with a greater mRNA decrease in tumours arising in younger patients. In contrast while there was only a trending interaction between age and mutation status on mRNA abundance (adjusted $p = 0.16$, **Figure 4D**), *ATRX* and age were synergistically associated with outcome, stratifying lower grade glioma patients into four groups (**Figure 4E**). While inactivating mutations in *ATRX* mutation are known to be generally associated with improved survival, younger patients without *ATRX* mutations have the best overall survival while high age patients without mutated *ATRX* have the worst survival, revealing its role as a strong age-dependent prognostic biomarker.

Discussion

Despite modest statistical power, suboptimal study designs and limited clinical annotation, we identified myriad age-associated differences in cancer genomes. Age-biased 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[38,64] , these data reveal a set of host influences on the mutational characteristics of tumours (**Figure 5**). Characteristics of the tumour host appears to influence all aspects of the cancer genome: mutation density, evolutionary timing, mutational processes and driver genes. Some of these lead to age-, ancestry- and sex-specific transcriptomic and clinical impacts.

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[65,66], somatic mosaicism and the accumulation of mutations over time[59,67,68]. However, other processes related to immune surveillance, evolutionary selection, disease aetiology and epigenetics are also likely involved[69–71]. In addition to such biological factors, lifestyle and socioeconomic considerations like diet[72] and microbiome composition[73] 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. A tumour's mutational history 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.

The TCGA and PCAWG datasets sometimes identified different molecular biases, highlighting the differences between the two datasets. TCGA patients were largely North American while PCAWG had a greater international component. While the ages represented in TCGA and PCAWG tumour-types were similar (**Table 1**), the cohorts differ in other host and clinical characteristics. For instance, the representation of ancestry groups was 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. Indeed, while we were able to identify more age-biases in TCGA data, many of these findings were reflected in PCAWG 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-biases 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[36,74,75]: 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 Yuan *et al*. was downloaded from The Cancer Genetic Ancestry Atlas (http://52.25.87.215/TCGAA).

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[37]. The individual data sets 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).

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 or insufficient variability in ancestry annotation were excluded from analysis. Age is treated as a continuous variable for both TCGA and PCAWG analyses. A full breakdown of the data is presented 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-biases of interest ($q < 0.1$). We used two-sided non-parametric univariate tests to minimize assumptions on the data. For putative age-biases, 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[76].

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 and PCAWG 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. 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.**

Genome instability

Performed for both TCGA and PCAWG 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.**

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-biases were identified using univariate logistic regression and putative biases were further analysed using multivariate logistic regression.. A multivariate q-value threshold of 0.1 was used to determine statistically significant age-biased 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-biased 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-biased 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 both 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-biases. Signatures with putative age-biases 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-biased 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 both TCGA and PCAWG data. Adjacent genes whose copy number profiles across patients were highly correlated (Pearson's $r > 95\%$) were binned. The copy number call for each patient was taken to be the majority call across all genes in each bin. 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**.

Genome-spanning SNV analysis

Performed for TCGA data. We focused on genes mutated in at least 1% of patients. Mutation data was binarized to indicate presence or absence of SNV in each gene per patient. Univariate logistic regression was used to identify putative age-biased SNVs. False discovery rate correction was used to adjust p-values and a q-value threshold of 0.1 used to select genes for multivariate analysis using logistic regression. SNV density was included in all multivariate models.

Driver Event Analysis

Performed for PCAWG data. We focused on driver events described by the PCAWG consortium[63]. 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-biased 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-biased 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-biased 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-biased gains, the same procedure

was applied using copy number gains. mRNA modeling results for age-biased 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-biased CNAs and SNVs were also analyzed using Cox proportional hazards modelling after checking proportional hazards assumption. Cox proportional hazard regression models incorporating ageest, 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-biased 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[77] (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. 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).
26. Willauer, A. N. *et al.* Clinical and molecular characterization of early-onset colorectal cancer. *Cancer* **125**, 2002–2010 (2019).
27. Lieu, C. H. *et al.* Comprehensive Genomic Landscapes in Early and Later Onset Colorectal Cancer. *Clin. Cancer Res.* **25**, 5852–5858 (2019).
28. Parsons, D. W. *et al.* An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science* **321**, 1807–1812 (2008).
29. Williams, G. R. *et al.* Comorbidity in older adults with cancer. *J. Geriatr. Oncol.* **7**, 249–257 (2016).

30. 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).
31. Balducci, L., Colloca, G., Cesari, M. & Gambassi, G. Assessment and treatment of elderly patients with cancer. *Surg. Oncol.* **19**, 117–123 (2010).
32. Given, B. & Given, C. W. Older adults and cancer treatment. *Cancer* **113**, 3505–3511 (2008).
33. 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).
34. 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).
35. Milholland, B., Auton, A., Suh, Y. & Vijg, J. Age-related somatic mutations in the cancer genome. *Oncotarget* **6**, 24627–24635 (2015).
36. Wahl, D. R. *et al.* Pan-Cancer Analysis of Genomic Sequencing Among the Elderly. *Int. J. Radiat. Oncol.* **98**, 726–732 (2017).
37. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
38. 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).
39. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, (2017).

40. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
41. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
42. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
43. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
44. Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
45. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
46. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).
47. 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).
48. 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).
49. Hieronymus, H. *et al.* Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11139–11144 (2014).
50. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).

51. Ratti, M., Lampis, A., Hahne, J. C., Passalacqua, R. & Valeri, N. Microsatellite instability in gastric cancer: molecular bases, clinical perspectives, and new treatment approaches. *Cell. Mol. Life Sci.* **75**, 4151–4162 (2018).
52. Boland, C. R. & Goel, A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* **138**, 2073-2087.e3 (2010).
53. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 15180 (2017).
54. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis. Oncol.* 1–15 (2017) doi:10.1200/PO.17.00073.
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. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
60. PCAWG Mutational Signatures Working Group *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
61. Alexandrov, L. *et al.* The Repertoire of Mutational Signatures in Human Cancer. (2018) doi:10.1101/322859.

62. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
63. 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).
64. Li, C. H. *et al. Sex Differences in Oncogenic Mutational Processes*. http://biorxiv.org/lookup/doi/10.1101/528968 (2019) doi:10.1101/528968.
65. 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).
66. Hoeijmakers, J. H. J. DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* **361**, 1475–1485 (2009).
67. Fernández, L. C., Torres, M. & Real, F. X. Somatic mosaicism: on the road to cancer. *Nat. Rev. Cancer* **16**, 43–55 (2016).
68. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
69. 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).
70. Cramer, D. W. & Finn, O. J. Epidemiologic perspective on immune-surveillance in cancer. *Curr. Opin. Immunol.* **23**, 265–271 (2011).
71. Vick, A. D. & Burris, H. H. Epigenetics and Health Disparities. *Curr. Epidemiol. Rep.* **4**, 31–37 (2017).
72. Research, W. C. R. F. I. for C. Diet, nutrition, physical activity and cancer: a global perspective. *Contin. Update Proj. Expert Rep.* (2018).

73. 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).

74. 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).

75. 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).

76. Yeo, I.-K. & Johnson, R. A. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* **87**, 954–959 (2000).

77. 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 discussions. This study was conducted with the support of the Ontario Institute for Cancer Research to P.C.B. through funding provided by the Government of Ontario. 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 (1U01CA214194-01).

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

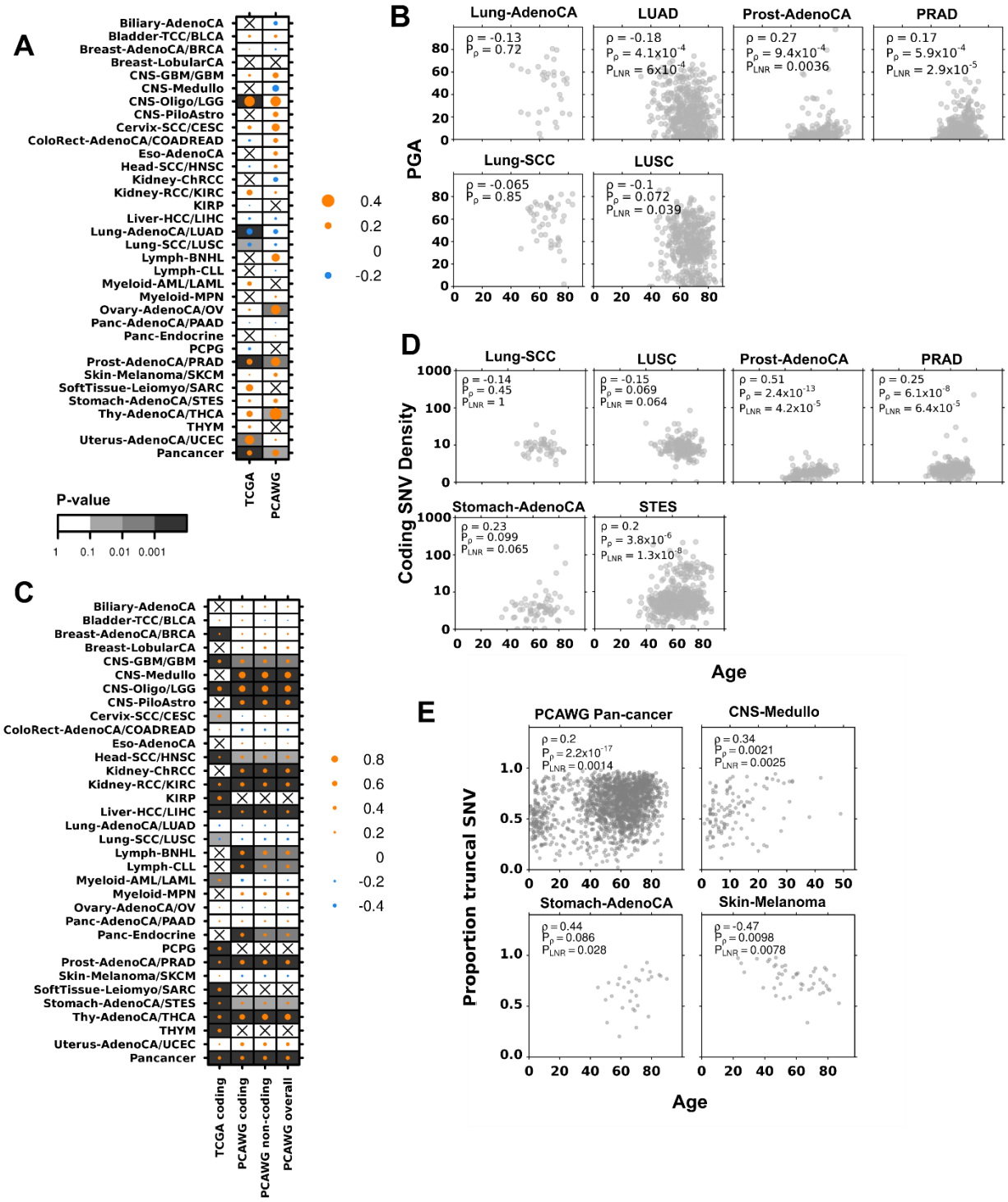

Figure 1 | Mutation density and timing are biased to age.
Summary of associations between age and **(A)** percent genome altered and **(C)** SNV density in TCGA and PCAWG tumours. The dot size and colour show the Spearman correlation, and background shading indicate adjusted multivariate p-value. Only tumour-types with at least univariately significant associations are shown. Associations between **(B)** PGA and **(D)** coding SNV density with age in selected tumour-type specific analyses. Univariate Spearman correlation, adjusted correlation p-value and adjusted multivariate p-values shown. **(E)** Correlations between age and proportion of SNVs occurring in the truncal clone in four PCAWG tumour contexts.

Figure 2

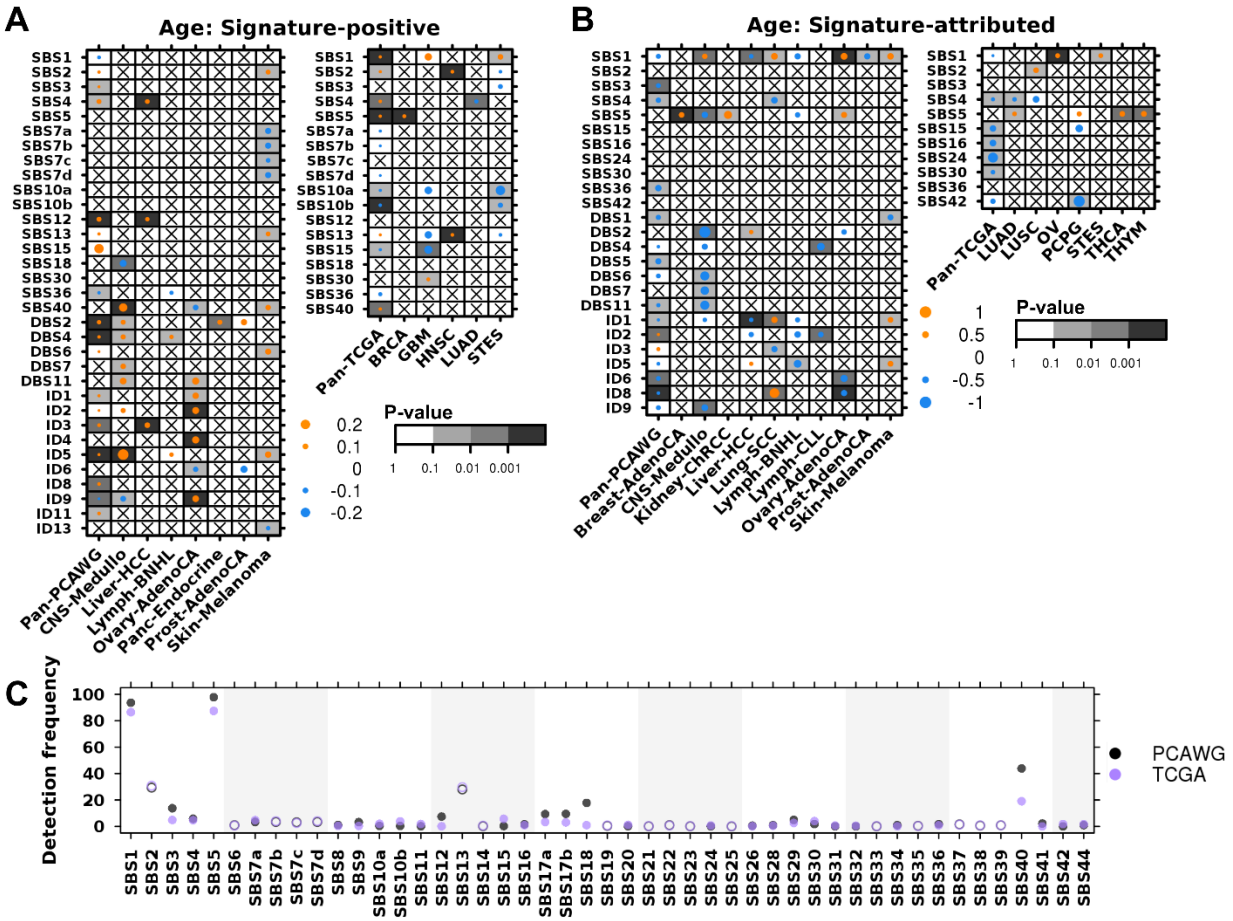

Figure 2 | Biases in mutational signatures suggest differences in underlying mutational processes.

(A) 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. **(B)** 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. **(C)** 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.

Figure 3

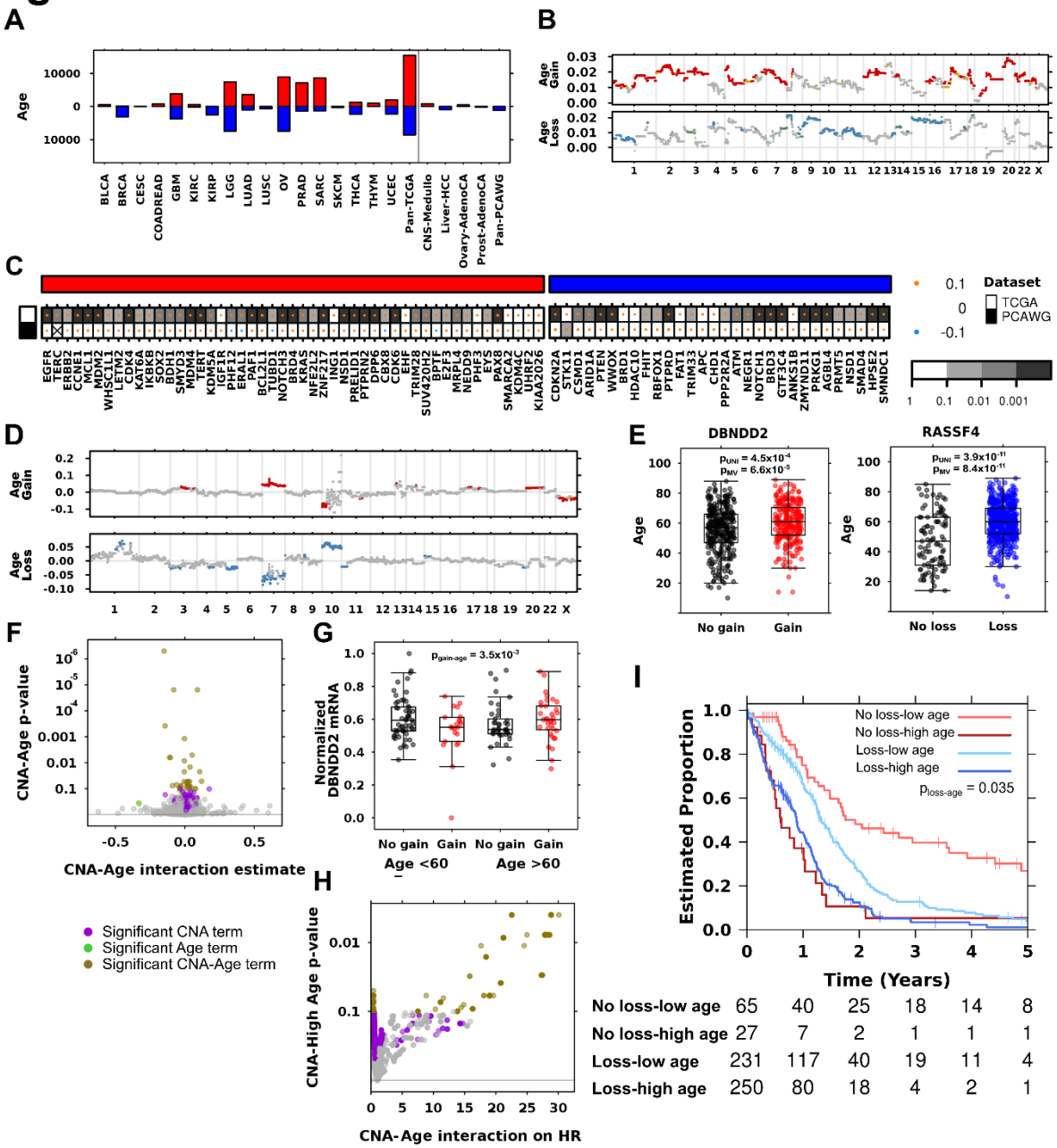

Figure 3 | Age-biases in copy number alterations are associated with functional changes in mRNA and survival.

(A) Summary of all detected age-biased CNAs with numbers of gains (above x-axis) and losses (below x-axis) found in each tumour context. Only tumour-types with at least one significant event shown. **(B)** Pan-cancer age-biases in CNAs for TCGA data. Each plot shows the logistic regression log odds coefficient estimate for the indicated CNA type. Dot colour indicates statistical significance, where red (copy number gain) and blue (copy number loss) show adjusted $p < 0.05$ and yellow (gain) and green (loss) show adjusted $p < 0.1$ **(C)** Summary of age-biased pan-cancer CNA drivers. Both TCGA and PCAWG

findings shown. Dot size shows the magnitude of the association as the difference in proportion and the background shading shows adjusted multivariate p-values. Top covariate indicates copy number gain drivers in red and loss drivers in blue. **(D)** Age-biases in TCGA glioblastoma CNAs across the genome with **(E)** specific examples shown for *DBNDD2* gain and *RASSF4* loss. Both adjusted univariate and multivariate p-values shown. Age-biased CNAs in TCGA glioblastoma are associated with **(F)** mRNA abundance changes and **(H)** overall survival. In **(F)**, adjusted p-values are plotted against the coefficients of the CNA-age interaction for mRNA abundance, with each point representing a gene. Dot colour shows significant associations between mRNA and age (green), CNA only (violet) or their interaction (olive). **(G)** *DBNDD2* mRNA abundance changes between copy number gain (red) or no loss (black) in tumours of low *vs*. high age. Adjusted CNA-age interaction p-value is shown. In **(H)**, adjusted p-values and coefficients of the CNA-age interaction for Cox-PH modeling are shown, with each point representing a gene. **(I)** Loss of a region on 10q interacts with age to further stratify patient prognosis. The adjusted p-value for the copy number loss-age interaction term 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.

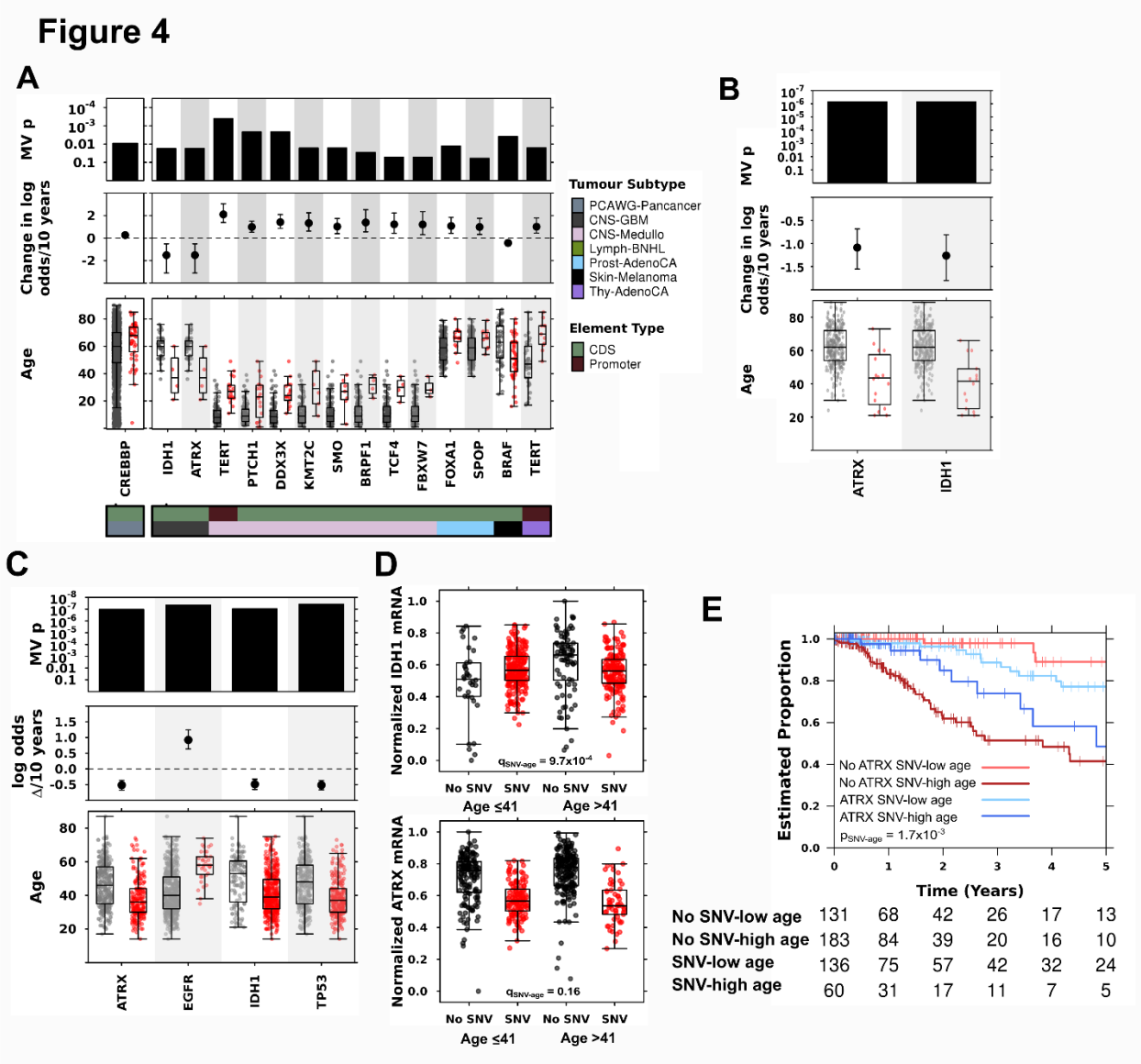

Figure 4 | Age-biases in single nucleotide variants reveal ATRX as a strong age-biased prognostic biomarker in lower grade glioma.

(A) Pan-PCAWG and PCAWG tumour-specific age-biases 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. Genes are ordered by tumour-type and adjusted p-value. Similar to **(A)**, genes with age-associated mutation frequency are shown for **(B)** TCGA glioblastoma and **(C)** TCGA lower grade glioma. **(D)** mRNA abundance changes for *IDH1* and *ATRX* when the gene is mutated (red) or not (black) compared by median-dichotomized age. Adjusted CNA-age interaction p-value is shown. **(E)** *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.

Figure 5

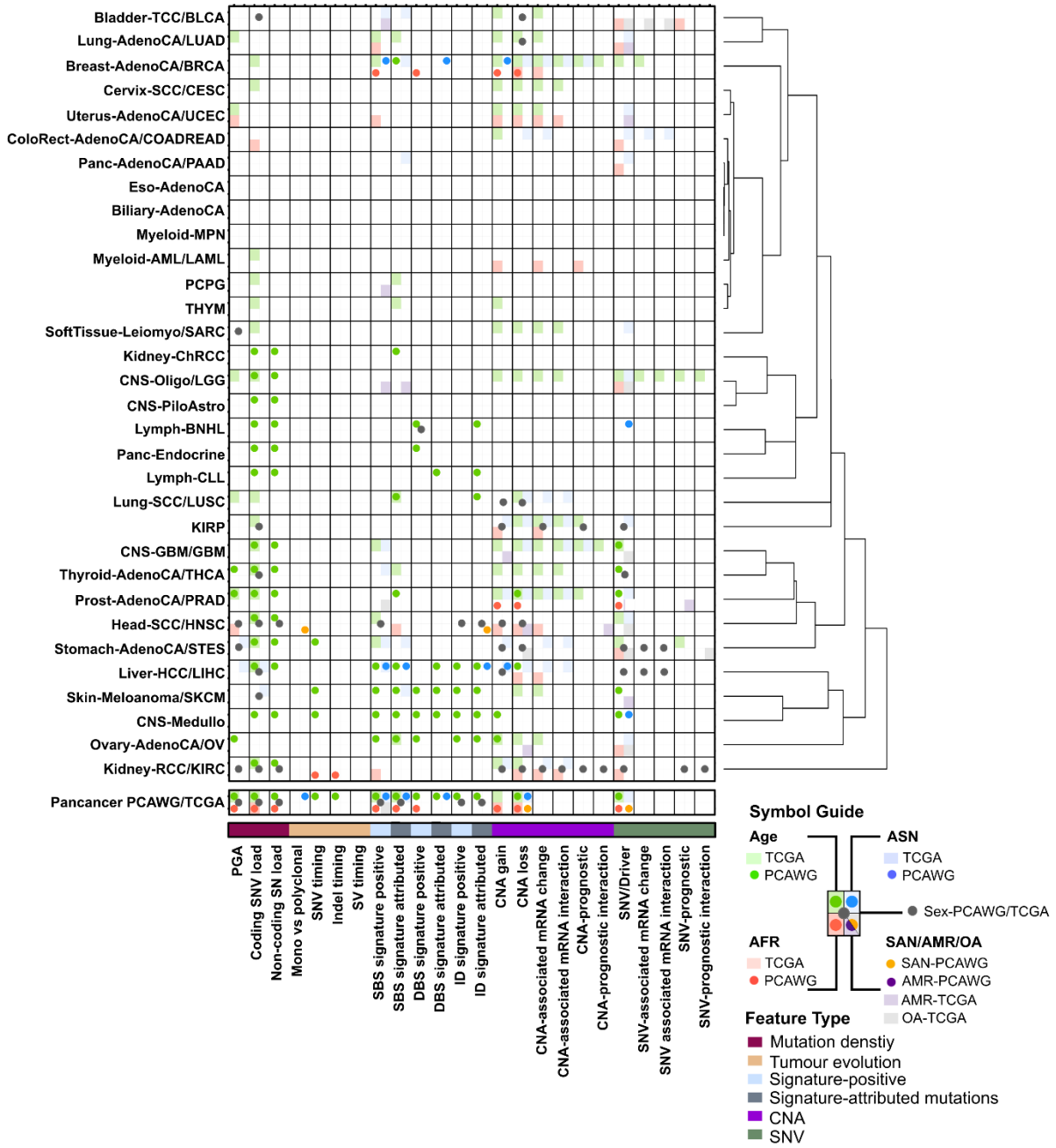

Figure 5 | The landscape of age, sex and ancestry differences in cancer genomics A summary of age-, ancestry- and sex-associated biases in TCGA and PCAWG analyses. Dots show associations in PCAWG data and shading shows associations in TCGA data. Each quadrant of every cell corresponds with age, Asian, African, and Admixed American, South Asian or Other Ancestry-associated findings. Centre dot indicates sex-associated findings.