1 **Somatic mutational profiles and germline polygenic risk scores in human cancer**

2 Yuxi Liu, MS^{1,2}, Alexander Gusev, PhD³, Yujing J. Heng, PhD⁴, Ludmil B. Alexandrov, PhD⁵, Peter

3 Kraft, PhD^{1,2,6,*}

4 1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

5 2. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of

6 Public Health, Boston, MA

7 3. Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical

8 School, Boston, MA

9 4. Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School,

10 Boston, MA

11 5. Department of Cellular and Molecular Medicine, UC San Diego, La Jolla, CA, 92093, USA

12 6. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

13 * Corresponding author: Peter Kraft, PhD. Program in Genetic Epidemiology and Statistical

14 Genetics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Building 2-249A,

15 Boston, MA, 02115. Email: pkraft@hsph.harvard.edu

16 **ABSTRACT**

17 The mutational profile of a cancer reflects the activity of the mutagenic processes which have
18 been operative throughout the lineage of the cancer cell. These processes leave characteristic
19 profiles of somatic mutations called mutational signatures. Mutational signatures, including
20 single-based substitution (SBS) signatures, may reflect the effects of exogenous or endogenous
21 exposures. Here, we used polygenic risk score (PRS) as proxies for exposures and examined the
22 association between somatic mutational profiles and germline PRS in 12 cancer types from The
23 Cancer Genome Atlas project. We found 17 statistically significant associations after Bonferroni
24 correction ($p < 3.15 \times 10^{-5}$), including positive associations between germline inflammatory
25 bowel disease PRS and number of somatic mutations of signature SBS1 in prostate cancer and
26 APOBEC-related signatures in breast cancer. The age at menarche PRS was inversely associated
27 with mutation counts of SBS1 in prostate cancer. Our analysis suggests that there are robust
28 associations between tumor somatic mutational profiles and germline PRS. These may reflect
29 mechanisms through hormone regulation and immunological responses that contribute to
30 cancer etiology and drive cancer progression.

INTRODUCTION

Cancer is driven by the accumulation of somatic mutations. In contrast to germline variants, which are inherited from egg or sperm and occur in DNA of every cell in the body, somatic mutations are generated from mutational processes of exogenous and endogenous exposures as well as DNA enzymatic modifications and failure/infidelity of DNA repair and replication^{1, 2, 3}. Mutational processes result in different mutation types (e.g., C>T substitution at the mutated base of ACG motif) with characteristic combinations of mutation types constituting different mutational signatures^{1, 4}. Previous studies have identified and confirmed more than 50 distinct signatures of single-base substitution (SBS) derived from the analysis of whole genome and whole exome sequences (WES) of multiple cancer types^{1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16}, but the etiologies for many of these signatures remain largely unexplored. In addition, tumor mutational burden (TMB), which quantifies the total mutations per megabase in a tumor tissue, has been suggested as a biomarker to predict the response of a patient to immunotherapy^{17, 18, 19}. Recently, The US Food and Drug Administration approved the use of pembrolizumab, a humanized antibody for cancer immunotherapy, in patients with TMB-High solid tumors. However, currently, it is not fully understood the reason most patients with high TMB benefit from immunotherapy.

Mutational signatures reflect the activity of the mutational processes that have been active through a person's life^{2, 20, 21, 22}. The identified mutational signatures reflect both processes commonly found across cancer types as well as processes confined to a particular cancer type. For example, signature SBS2 and SBS13, both attributed to the enzymatic activity of the

APOBEC family of cytidine deaminases, are present in multiple cancer types⁶. In contrast, signature SBS12, whose etiology is still unknown, is almost exclusively found in liver cancers¹. Some signatures reflect the effects of lifestyle choices, such as signature SBS4, which is associated with tobacco smoking in multiple cancer types, or environmental exposures, such as signatures SBS7a/b/c/d which are imprinted by exposure to ultraviolet light^{1, 23}. Some signatures are caused by endogenous exposures, for example, the clock-like signature SBS1 is attributed to endogenous deamination of 5-methylcytosine²⁴. However, the etiologies for many other recently identified signatures remain unclear. Linking those signatures of unknown origin to cancer risk factors may suggest mechanism and provide avenues for further investigation.

Previous studies of multiple cancer types have found associations between somatic mutational burden and germline genetic risk factors²⁵. For example, germline *MC1R* R alleles carrier status is significantly associated with somatic mutational burden in melanomas²⁶. Germline statuses of *ZNF750* and *CDC27* have an impact on somatic mutational signatures in esophageal squamous cell carcinomas²⁷. rs2588809 carrier status at gene *RAD51B* is significantly associated with TMB in breast cancer²⁸. rs17000526, a variant associated with *APOBEC3B* expression, is strongly associated with APOBEC signature mutations in bladder cancer²⁹. Carter et al.³⁰ investigated the interaction between an array of germline variants and somatic events across cancers and found robust associations. In addition to studying the association at the level of individual variants, one study also examined the relationship between polygenic risk score (PRS), which combines the effect of multiple germline variants, and somatic mutational burden. They found that germline PRS of breast cancer was inversely associated with total somatic mutation counts in breast tumor samples²⁸, but the underlying mechanism is still obscure.

74 Here, we performed a pan-cancer analysis of the association between tumor somatic
75 mutational profiles and germline PRS of cancers and non-cancer traits using data from The
76 Cancer Genome Atlas (TCGA). Studies with comprehensive somatic mutation data do not
77 always have complete and accurate epidemiological exposure data. PRS can be used as
78 instruments for those unmeasured exogenous and endogenous risk factors. Studying the
79 relationship between germline PRS and somatic mutations can also provide insight into the
80 underlying biological mechanisms of cancer development.

RESULTS

TCGA germline and somatic data. We obtained tumor somatic mutation data, germline variant data, and clinical data for 5,766 patients of European ancestry across 12 cancer types from TCGA. The sample size for each cancer type ranged from 333 (kidney renal clear cell carcinoma) to 929 (breast invasive carcinoma). A mutational signature was included in the analysis if there was at least one cancer type where both more than 20 samples and more than 20% of the samples in the cancer type had mutations attributed to that signature. Overall, somatic data were retrieved for (Fig. 1): (i) total somatic mutation counts (TSMC), defined as the total number of somatic missense mutations³¹; (ii) nine individual SBS mutational signatures; and (iii) one combined APOBEC-related signatures (SBS2 and SBS13).

Calculation of germline PRS. We calculated germline PRS for 23 cancers and non-cancer traits, including the 12 selected cancer types, breast cancer stratified into estrogen receptor positive (ER+) and negative (ER-) subtypes, cancer risk determinants (cigarettes per day, drink per week, and body mass index (BMI)), hormonal factors (age at menarche and age at natural menopause), and immune-mediated inflammatory diseases (inflammatory bowel disease (IBD), ulcerative colitis (UC), Crohn's disease (CD), and rheumatoid arthritis (RA)) using the TCGA germline variant data and published summary statistics from genome-wide association studies (GWAS) (see Methods, Supplementary Table 1). The PRS validation results are summarized in Supplementary Table 2 (see Methods): all cancer PRS were positively associated with the corresponding cancer in the TCGA sample.

Correlations with age at cancer diagnosis. We checked the correlations between somatic mutations by signature (and TSMC) and age at cancer diagnosis for each selected cancer type (Fig. 2). Consistent with previous study⁵, SBS1 and SBS5, the two clock-like signatures for which the numbers of mutations increase with age, showed strong positive correlations with age at diagnosis for most cancer types, though there exists heterogeneity across cancers. We also evaluated the correlation between the calculated germline PRS and age at diagnosis for each cancer type (Supplementary Fig. 1). Although none of these associations passed the Bonferroni-adjusted statistical significance threshold accounting for 322 tests ($p = 1.55 \times 10^{-4}$), the Spearman's ρ with age for most cancer PRS were negative among the cases of that specific cancer, indicating that higher germline PRS of a cancer is associated with earlier diagnosis of that cancer.

Associations between somatic mutations and germline PRS. We fit a zero-inflated negative binomial model, negative binomial model, or linear model regressing tumor somatic mutation counts on germline cancer or non-cancer PRS for each combination of cancer type (or subtype), SBS signature (or TSMC), and PRS, adjusting for sex (if applicable), age at cancer diagnosis (in years), and the top 10 genetic principal components (PCs). The p-value threshold for significance was 3.15×10^{-5} , adjusting for multiple comparisons using Bonferroni correction (1,587 tests). We found 17 statistically significant associations (Table 1). Significant associations were found for prostate, breast, and endometrial cancers; most of them involved somatic mutation count of SBS1 (deamination of 5-methylcytosine) and the immune-mediated inflammatory diseases (IBD, CD, and UC) PRS. The association results across the 12 cancer types and two breast cancer subtypes for significant signature-PRS pairs are shown in Supplementary

Fig. 2. Given that there was a strong positive correlation between number of mutations of SBS1 and age in breast cancer (Spearman's $\rho = 0.18$, $p = 2.50 \times 10^{-6}$), a nominally significant correlation between SBS1 and age in prostate cancer (Spearman's $\rho = 0.12$, $p = 0.02$), and a significant correlation between SBS40 and age in endometrial cancer (Spearman's $\rho = 0.18$, $p = 9.67 \times 10^{-4}$), we further evaluated the associations of somatic mutations and PRS without adjusting for age at cancer diagnosis. There was no substantial change on the top findings, though we observed more significant associations with SBS1 and SBS5, which are likely to be driven by their correlations with age at diagnosis (Supplementary Table 3). The association between APOBEC-related signatures and breast cancer PRS in breast tumor is slightly above the Bonferroni threshold ($p = 4.21 \times 10^{-5}$) without adjusting for age.

There was no substantial change in the directions and effect sizes of the significant associations after adjusting for tumor stage where available (see Methods); the p-values for the breast cancer associations remained nominally significant after adjusting for stage (Table 1).

Consistent with a previous study²⁸, which found a significant inverse association between breast cancer PRS and TSMC in breast tumor samples for ER+ patients, we found inverse associations between the somatic mutation counts of APOBEC-related signatures and breast cancer PRS (both overall and ER+ specific) in breast tumors. We further adjusted for the known germline APOBEC3 risk variant, rs17000526²⁹, in the associations with the APOBEC-related signatures. This variant was not included in the calculation of any cancer or non-cancer PRS or in linkage disequilibrium (LD) ($r^2 > 0.1$) with any PRS variants. There was no substantial change on the top associations with the APOBEC-related signatures. Consistent with a previous study²⁹,

we observed a significant association between rs17000526 and the APOBEC-related signatures in bladder tumors (regression coefficient = 0.21 for the rs17000526-A allele, $p = 1.81 \times 10^{-4}$).

This association was not significant in overall breast tumors. However, we did find a nominally significant association between rs17000526 and the APOBEC-related signatures in ER+ breast tumors (regression coefficient for the A allele = 0.10, $p = 0.04$).

Except for breast cancer, none of the somatic mutations and germline PRS associations for the same cancer type passed the significance threshold; the smallest p-value observed was for the positive association between prostate cancer PRS and SBS1 count in prostate tumor ($p = 1.12 \times 10^{-3}$). We did not see any significant association between mutation counts in lung tumor and cigarettes-per-day PRS (smallest p-values: $p = 6.22 \times 10^{-3}$ for TSMC in lung adenocarcinoma; $p = 0.02$ for SBS1 in lung squamous cell carcinoma). The smallest p-value for the cigarettes-per-day PRS was observed for SBS1 count in colorectal tumor ($p = 1.70 \times 10^{-4}$). Drinks-per-week and RA PRS were also not significantly associated with the mutation counts in any cancer type (smallest p-values: $p = 2.26 \times 10^{-3}$ for drinks-per-week PRS and SBS1 count in lung adenocarcinoma; $p = 8.21 \times 10^{-4}$ for RA and APOBEC-related signatures count in ER- breast tumor).

We further performed a meta-analysis on the association of SBS signatures (or TSMC) and germline PRS across cancers. The associations of SBS1 with CD PRS, SBS1 with bladder cancer PRS, and APOBEC-related signatures with IBD, CD, and UC PRS were significant using both fixed-effect and the Stouffer's Z methods ($p < 4.35 \times 10^{-4}$ after Bonferroni correction for 115 tests) (Fig. 3). Overall, we found 17 associations that were statistically significant based on either the fixed-effect model or Stouffer's method (Supplementary Table 4).

DISCUSSION

We performed a comprehensive analysis on the association between tumor somatic mutational profiles and germline PRS for various cancers and non-cancer traits, leveraging mutational signatures and germline variant data of 12 cancer types, as well as summary statistics from recent large GWAS. Our results demonstrate that there are robust associations between somatic mutational burden and germline PRS in human cancer. Some PRS can serve as proxies for exposures: linking mutational signatures to these exogenous and endogenous risk factors represented by PRS may suggest the etiology of associated mutational processes. Other PRS reflect the germline genetic contribution to cancer risk: studying their relationships with tumor somatic mutational burden may shed light on the underlying mechanism of cancer development that attributed to germline-somatic interactions.

For cancer PRS, the overall breast cancer PRS and ER+ specific PRS were both associated with the somatic mutation count of APOBEC-related signatures in breast tumor. These inverse relationships are consistent with a previous study by Zhu et al.²⁸. Similarly, Qing et al.³² reported a significant inverse relationship between rare, high impact germline coding variants and somatic mutational burden in cancer hallmark genes among TCGA patients across age groups. It has been hypothesized that patients with higher germline variant burden tend to develop cancer at a younger age thus would have fewer acquired somatic mutations, whereas patients at a lower germline genetic risk may need longer time for cancer development, which is mainly driven by the accumulation of somatic mutations^{28, 32}. However, we observed this inverse relationship for APOBEC-related signatures, which did not show clock-like behavior in

previous studies^{1, 5}. In our data, there was also no statistically significant correlation between mutation count of APOBEC-related signatures in breast cancer and age at diagnosis (Spearman's $\rho = -0.01$, $p = 0.73$, Fig. 2). The p-value for this association increased after adjusting for tumor stage, but the direction of association did not change. This inverse relationship may indicate that breast tumors in women with low breast cancer PRS are predisposed to develop tumors with higher APOBEC mutation counts (due to yet-to-be-determined biological mechanisms), but it may also be a result of collider bias (Fig. 3a): if other breast cancer risk factors that are independent of the breast cancer PRS in the general population are associated with higher APOBEC mutation counts, then the breast cancer PRS and APOBEC mutation counts could be negatively correlated among cases. Few studies have directly assessed the associations between established cancer risk factors and SBS signatures. A recent study reported no impact of BMI, cigarette smoking, and alcohol consumption on APOBEC signatures in TCGA³³.

In another simple setting assuming no reverse causation, there might be three possible causal relationships between cancer or non-cancer PRS, mutational signature, and cancer diagnosis given that a significant association between a germline PRS and a mutational signature is observed (Fig. 3b). Collider bias would also arise if both the PRS and mutational signature have a direct effect (or effect through other pathways) on tumor development and cancer diagnosis. If that is the case, a non-causal association between germline PRS and mutational signature would be observed if we only include cancer cases in the study. This scenario is plausible for breast cancer, as many breast cancer susceptibility variants identified from GWAS have been shown to directly code or regulate cancer susceptibility genes (e.g., missense variant rs35383942 is in *PHLDA3*, which encodes a p53-regulated repressor of Akt³⁴), thus have effects

on breast cancer development through other functional pathways that are not mediated by the
 acquirement of somatic mutations^{35, 36, 37, 38}. To avoid the potential collider biases, somatic
 mutation profiling needs to be performed on both tumor and normal tissue at the cancer sites
 of interest and the sampling of study population should be independent of cancer status. The
 Human Tumor Atlas Network³⁹, which seeks to construct comprehensive atlases of molecular
 and cellular features of cancers, starting from precancerous lesions to advanced disease, could
 address these questions. In addition to the breast cancer associations, inverse relationships
 were also observed for other two cancer PRS associations (head and neck cancer PRS and
 glioblastoma PRS with SBS1 in prostate tumor), either adjusting or not adjusting for age at
 diagnosis, which may also be explained by the collider bias described above.

For non-cancer PRS, germline PRS of age at menarche was found to be significantly associated
 with SBS1 count in prostate tumor. Specifically, we found that prostate cancer patients with
 higher age at menarche PRS, a surrogate for later menarche, tend to have fewer SBS1
 mutations. Although males do not exhibit menarche, a previous study reported a strong genetic
 correlation ($r_g = 0.74$) between female and male puberty timing, represented by age at
 menarche and age at voice breaking, respectively⁴⁰. Therefore, we hypothesize that the effect
 of hormonal factor PRS on the number of somatic mutations in prostate tumors may be
 explained by the shared regulatory mechanism of puberty timing in men and women. Prostate
 cancer has long been recognized as a hormone-related cancer, traditional epidemiological
 studies have reported associations between various hormone-related markers (e.g., Insulin-like
 growth factor 1 (IGF-1), testosterone, sex hormone-binding globulin, etc.) and the risk of
 prostate cancer^{41, 42, 43, 44}. Mendelian randomization studies also reported a consistent

protective effect of later pubertal development on prostate cancer risk^{40, 45}. All these suggest an etiological relevance between the timing of puberty and incidence of prostate cancer involving shared effect of hormonal factors. Previous findings highlighted the roles of androgen and IGF-1, of which the circulating levels increase dramatically during puberty, in prostate carcinogenesis^{45, 46, 47, 48, 49}. They proposed that the concentrations of these hormones during this susceptibility window when the luminal cells start to appear and the prostate becomes mature may have an impact on prostate cancer risk in later life^{50, 51, 52, 53}. Our finding suggests a potential pathway through the hormone-related markers of puberty timing on the accumulation of SBS1 mutations in prostate tissue, which may drive tumor development.

Inflammatory bowel disease PRS were found to be associated with somatic mutational profiles in multiple hormone-related cancers (breast, endometrial, and prostate cancer). Previous studies have established CD and UC as risk factors for overall cancer^{54, 55, 56, 57, 58}, but whether these associations are driven by shared genetic susceptibility or other common lifestyle and environmental factors remains unanswered by these epidemiological studies. In the present work, we found robust positive associations between IBD PRS (IBD, UC, and CD) and somatic mutations in breast cancer. Previous evidence has shown an increased risk of breast cancer among UC and CD patients, and first-degree relatives of CD patients^{57, 59, 60}. Several potential mechanisms have been proposed. Interleukin-1 (IL-1) polymorphism has been linked the risk of many diseases, including breast cancer and IBD^{61, 62, 63, 64, 65, 66}. A recent study⁶⁷ identified 53 common differentially expressed genes for breast cancer and CD compared with controls, three major hub genes, *CXCL8*, *IL-1β* and *PTGS2*, and the shared IL-17 and NF-κB signaling pathways. These suggest that the elevation of cytokines may play a common role in the pathogenesis of

CD and breast cancer. Hovde et al.⁵⁷ also proposed that the downregulation of breast cancer resistance protein in UC patients may be related to breast cancer etiology. In addition to breast cancer, we also observed positive relationships between IBD PRS (IBD, UC, and CD) and somatic mutation count of SBS1 in prostate tumor. Moreover, the association between CD and SBS1 is broadly significant across the 12 cancer types in our study. Recent meta-analyses and a large cohort study all conclude that patients with IBD have an increased risk of prostate cancer^{68, 69, 70}, though the previous findings were inconclusive^{71, 72, 73, 74, 75, 76}. Previous studies have proposed that shared risk alleles may partially explain this association^{68, 75, 77}. Folate hydrolase 1 (FOLH1) or prostate-specific membrane antigen (PSMA) is highly expressed in both IBD and prostate cancer^{78, 79, 80}. Studies have reported that Inhibition of FOLH1/PSMA ameliorates IBD symptom in mice models⁷⁸ and also leads to tumor regression in preclinical models⁸¹. Our results suggest there is a link between inherited genetic variants that contribute to the development of inflammatory bowel disease and acquired somatic mutations in these hormonal-related cancers. Future studies need to further investigate the mechanism underlying the associations between specific PRS and specific SBS signatures, especially between CD and SBS1. One direction might be looking at the associations between these mutational signatures and the measured immune-related markers in TCGA. This may also provide novel insight into the mechanism underlying the association between TMB and benefit from immunotherapy. Interestingly, we found a significant inverse association between IBD PRS (IBD and CD) and SBS40 count in endometrial cancer. There are limited studies looking at the association between IBD and risk of endometrial cancer. A meta-analysis reported an inverse relationship,

but it is not statistically significant⁷¹. Our results may suggest potential mechanisms underlying SBS40 through immune-related processes.

A significant positive association was observed between BMI PRS and SBS1 count in prostate cancer. Obesity has been established as an independent risk factor associated with the risk of advanced or fatal prostate cancer^{82, 83, 84}. Studies also have shown that high BMI is associated with a decreased risk of low-grade prostate cancer^{82, 83, 85, 86, 87}. However, these associations may be due to diagnostic bias. Men with higher BMI tend to have lower prostate-specific antigen levels^{88, 89}, which make them less likely to have a biopsy⁹⁰, and larger prostate volumes, which makes it harder to find the tumor from biopsy⁹¹. Mendelian randomization studies have found little evidence for a causal relationship between BMI and prostate cancer incidence^{92, 93}. Therefore, it is likely that higher BMI PRS is associated with older age at diagnosis, and since SBS1 counts increase with age, people with higher BMI PRS would have higher SBS1 counts as well. We did not adjust for tumor stage or grade for prostate cancer cases, as it was unavailable. We adjusted for age at diagnosis, and the BMI PRS is not significantly associated with age in prostate cancers in our study (Supplementary Fig. 1), but we still observed this strong positive association. Future studies need to further investigate this association with SBS1 incorporating tumor stage and grade information, and carefully control for the potential biases. Our study has limitations. The calculated PRS only explains a small proportion of the heritability thus may not fully represent the genetic variant burden underlying the trait. Although we adjusted for tumor stage for most cancer types, we did not adjust for or perform subgroup analysis by tumor grade, as this information is not available for the selected TCGA patients.

Bonferroni correction for multiple testing is conservative. We expect to identify more significant associations with a larger sample size. There are several strengths of our study. We performed a comprehensive pan-cancer analysis on the relationship between 14 cancer PRS and 9 non-cancer PRS, and TSMC as well as the number of mutations attributed to 10 SBS signatures across 12 cancer types. We used PRS as genetic instruments for endogenous and exogenous exposures to avoid bias from confounding and reverse causation in traditional observational studies. Also, we have sufficient sample size for detecting associations of similar strength as previous study²⁸ in each cancer type with a high power, which may not be able to achieve using exposure data that usually have missing metadata. We controlled for the impact of age at diagnosis and tumor stage on the germline-somatic associations. Although PRS serves as a good proxy for exposure, it would still be useful to collect epidemiological data on exposures in future studies to confirm our findings and to better understand the role of germline variant burden underlying these associations. Future studies can also look at the association between signature-specific mutation count and immune features, hormone-related markers, and expression levels of cancer susceptibility genes to further investigate the underlying mechanisms.

In conclusion, our findings indicate that there are robust associations between somatic mutational profiles and germline PRS in human cancer. Our results demonstrate evidence for germline-somatic associations between inflammatory bowel disease PRS and somatic mutations (SBS1, APOEC-related signatures, SBS40, and TSMC) in breast cancer, endometrial cancer, and prostate cancer, and between age at menarche PRS and somatic mutations (SBS1)

315 in prostate cancer. Our results are relevant to the etiology of mutational signatures and the
316 underlying biological mechanisms of cancer development.

METHODS

Study population. We selected cancer types based on total number of cases in TCGA and the availability of large GWAS for calculating PRS. Twelve cancer types with WES data available on more than 300 primary tumor samples were selected to ensure 80% power (at a type I error rate of 5%) to detect an association between PRS and TSMC of similar or greater magnitude as that previously reported between rs2588809 and TMB²⁸ (Supplementary Fig. 3). Data were obtained for all TCGA patients of European ancestry from these 12 cancer types. We further restricted sample type to primary tumor and excluded male cases for breast cancer. Age at diagnosis, sex, and tumor stage were retrieved from the GDC data portal <https://portal.gdc.cancer.gov/>. Tumor stage was not available for prostate cancer, endometrial cancer, lower grade glioma, glioblastoma, and ovarian cancer. Samples with missing age at diagnosis, tumor stage (if available), and sex information were further excluded.

Mutational signatures. Mutational signatures were identified from TCGA WES data using methods based on nonnegative matrix factorization¹. We obtained the mutation counts for 54 distinct SBS signatures for each selected tumor sample from the ICGC data portal <https://dcc.icgc.org/releases/PCAWG>. In addition, we created a new signature variable called APOBEC-related signatures by summing up the somatic mutation counts of SBS2 and SBS13 (both attributed to APOBEC activity). For each cancer type, we analyzed those SBS signatures if there was at least one cancer type where both more than 20 samples and more than 20% of the samples in the cancer type had mutations attributed to that signature. In addition to the

SBS signature-specific mutations, we also retrieved TSMC for each sample using maftools R package⁹⁴.

Germline variant data. Raw germline single nucleotide polymorphism (SNP) array data were downloaded from the GDC Legacy Archive <https://portal.gdc.cancer.gov/legacy-archive>. Genotype quality control (QC) was applied to remove variants with >5% missing calls, or Hardy-Weinberg equilibrium $p < 5 \times 10^{-6}$, or minor allele frequency (MAF) <1%. To restrict to individuals of European ancestry, genetic PCs were computed using post-QC, LD pruned variants and any outlier sample >6 standard deviations away from the mean along either of the top 2 PCs was removed. All remaining samples and genotypes were then imputed to the Haplotype Reference Consortium reference panel⁹⁵.

Calculation of PRS. The PRS of a trait for subject i was calculated as,

$$PRS_i = \sum_j \beta_j G_{ij}$$

where the weight β_j was the log odds ratio (or the beta coefficient for continuous traits) of the trait comparing effect allele to other allele at SNP j , and G_{ij} was the expected number of effect allele at SNP j for subject i (allele dosage). The lists of SNPs for constructing PRS were obtained from published GWAS summary statistics (Supplementary Table 1).

For cancer PRS, the list of SNPs and corresponding weights were obtained from one of the three sources: (i) GWAS or PRS paper; (ii) NHGRI-EBI GWAS Catalog⁹⁶; and (iii) Cancer PRSweb⁹⁷. For non-cancer PRS, the summary statistics were all from GWAS or PRS papers. Sources of GWAS summary statistics for each trait are summarized in Supplementary Table 1.

We filtered the SNP list for each trait using different strategies. For those from GWAS Catalog, we removed results from cross-cancer, subgroup, or interaction analysis, and restricted to European ancestry studies with a minimum of 2,000 cases and 2,000 controls. We removed SNPs with p-value above the genome-wide significance threshold ($p > 5 \times 10^{-8}$) for those from GWAS Catalog or GWAS papers. For SNPs from Cancer PRSweb, we used the p-values threshold with the best performance as evaluated by Nagelkerke's pseudo- R^2 . We did not additionally filter by p-values nor perform LD clumping on SNPs from PRS papers. LD clumping was performed on SNP lists from GWAS paper and GWAS Catalog: we removed SNPs with $MAF < 1\%$ or in LD ($r^2 > 0.1$) with SNPs of smaller p-value. We used PLINK 2.0⁹⁸ to calculate PRS for each trait and subject from the final list of SNPs.

Validation of PRS. We evaluated the ability of the calculated PRS to discriminate between specific cancer cases and controls. For each cancer type and PRS, an unadjusted logistic model was fit by treating all patients of that cancer type as cases and a randomly selected subset (with the same size as cases) of other cancer cases as controls. The performance of PRS was assessed by area under the receiver operating characteristic curve (Supplementary Table 2). In addition to the 12 cancer types and two subtypes, we also evaluated the performance of PRS on discriminating lung cancer cases combined (lung adenocarcinoma and squamous cell carcinoma) and glioma cases combined (lower grade glioma and glioblastoma). For those cancer types with subtypes or other closely related types (e.g., breast cancer), none of the patients of the other subtypes (or related types) was included as controls when treating one subtype (or type) as cases.

Association between somatic mutation count and PRS. TSMC were Winsorized to 98%, where the counts below the first percentile were set to the first percentile and the counts above the 99th percentile were set to the 99th percentile. PRS was standardized to have mean zero and unit standard deviation before running each model. For each cancer type and SBS signature (or TSMC), we first fit a zero-inflated negative binomial model if the proportion of zero-count samples is greater than 10%, otherwise we fit a negative binomial model directly. If the zero-inflated negative binomial model failed to converge and the proportion of zero-count samples is less than 50% (if greater than or equal to 50%, all models were considered as failed), we tried negative binomial model. If both two models failed, we transformed the mutation count to $\log_{10}(\text{count} + 1)$ and fit the linear model. In each attempt, we ran both model with and without PRS (models were considered as failed if either model failed) and obtained p-value from likelihood ratio test on these two models. We performed a fixed-effect meta-analysis of the associations between TSMC or each SBS signature and each PRS across cancers. Stouffer's Z-score method was also used to combining the individual p-values from each cancer type. The Z-score for the overall meta-analysis is:

$$Z \sim \frac{\sum_{i=1}^k w_i Z_i}{\sqrt{\sum_{i=1}^k w_i^2}}$$

where Z_i is the Z-score for cancer type i , w_i is the sample size of cancer type i , k is the total number of cancer types and subtypes. We used Bonferroni correction to adjust for multiple comparisons of 115 tests, which is the total number of signature-PRS pairs, excluding those only

available for breast cancer (overall, ER+, and ER-) and lung cancer (adenocarcinoma and squamous cell carcinoma).

Sensitivity analysis. To assess the impact of age at cancer diagnosis on the association of germline PRS and somatic mutational burden, we ran the association tests without adjusting for age at diagnosis (same model types as adjusting for age). Spearman correlations were calculated for (i) somatic mutation count and age at diagnosis for each cancer type and signature. We used Bonferroni correction accounting for 69 tests, which is the total number of cancer-signature pairs included in the analyses; and (ii) PRS and age at diagnosis for each cancer type. We used Bonferroni correction accounting for 322 tests, which is the total number of cancer-PRS pairs. Tumor stage is only available for some cancers; we re-ran the association tests adjusting for tumor stage for those cancer types (same model types as not adjusting for tumor stage). We further adjusted for the known germline APOBEC3 risk variant, rs17000526, in the associations with the APOBEC-related signatures (same model types as not adjusting for rs17000526).

DATA AVAILABILITY

The TCGA germline variants data are available at <https://portal.gdc.cancer.gov/legacy-archive>. The TCGA somatic mutations and mutational signatures data are available from the maftools R package⁹⁴ and at <https://dcc.icgc.org/releases/PCAWG>. The TCGA clinical data are available at <https://portal.gdc.cancer.gov/>. Other relevant data are available from the authors upon request.

417 REFERENCES

- 418 1. Alexandrov LB, *et al.* The repertoire of mutational signatures in human cancer. *Nature*
419 **578**, 94-101 (2020).
420
- 421 2. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in
422 human cancers. *Nat Rev Genet* **15**, 585-598 (2014).
423
- 424 3. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* **458**, 719-724 (2009).
425
- 426 4. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering
427 signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-259
428 (2013).
429
- 430 5. Alexandrov LB, *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet*
431 **47**, 1402-1407 (2015).
432
- 433 6. Alexandrov LB, *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
434 415-421 (2013).
435
- 436 7. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93
437 (2020).
438
- 439 8. Hayward NK, *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature*
440 **545**, 175-180 (2017).
441
- 442 9. Kim J, *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature
443 in urothelial tumors. *Nat Genet* **48**, 600-606 (2016).
444
- 445 10. Merlevede J, *et al.* Mutation allele burden remains unchanged in chronic
446 myelomonocytic leukaemia responding to hypomethylating agents. *Nat Commun* **7**,
447 10767 (2016).
448
- 449 11. Nik-Zainal S, *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell*
450 **149**, 979-993 (2012).
451
- 452 12. Nik-Zainal S, *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome
453 sequences. *Nature* **534**, 47-54 (2016).
454
- 455 13. Polak P, *et al.* A mutational signature reveals alterations underlying deficient
456 homologous recombination repair in breast cancer. *Nat Genet* **49**, 1476-1486 (2017).
457

14. Poon SL, *et al.* Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med* **7**, 38 (2015).
15. Schulze K, *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet* **47**, 505-511 (2015).
16. Tate JG, *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
17. Chan TA, *et al.* Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann Oncol* **30**, 44-56 (2019).
18. Fancello L, Gandini S, Pelicci PG, Mazzarella L. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *J Immunother Cancer* **7**, 183 (2019).
19. Goodman AM, *et al.* Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* **16**, 2598-2608 (2017).
20. Bolli N, *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).
21. Bolli N, *et al.* Genomic patterns of progression in smoldering multiple myeloma. *Nat Commun* **9**, 3363 (2018).
22. Kasar S, *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* **6**, 8866 (2015).
23. Pfeifer GP. Environmental exposures and mutational patterns of cancer genomes. *Genome Med* **2**, 54 (2010).
24. Pfeifer GP. Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol* **301**, 259-281 (2006).
25. Ramroop JR, Gerber MM, Toland AE. Germline Variants Impact Somatic Events during Tumorigenesis. *Trends Genet* **35**, 515-526 (2019).
26. Robles-Espinoza CD, *et al.* Germline MC1R status influences somatic mutation burden in melanoma. *Nat Commun* **7**, 12064 (2016).
27. Guo J, *et al.* Germline and somatic variations influence the somatic mutational signatures of esophageal squamous cell carcinomas in a Chinese population. *BMC Genomics* **19**, 538 (2018).

28. Zhu B, *et al.* An investigation of the association of genetic susceptibility risk with somatic mutation burden in breast cancer. *Br J Cancer* **115**, 752-760 (2016).
29. Middlebrooks CD, *et al.* Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat Genet* **48**, 1330-1338 (2016).
30. Carter H, *et al.* Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov* **7**, 410-423 (2017).
31. Carbone DP, *et al.* First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2415-2426 (2017).
32. Qing T, *et al.* Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. *Nat Commun* **11**, 2438 (2020).
33. Heng YJ, *et al.* The Association of Modifiable Breast Cancer Risk Factors and Somatic Genomic Alterations in Breast Tumors: The Cancer Genome Atlas Network. *Cancer Epidemiol Biomarkers Prev* **29**, 599-605 (2020).
34. Kawase T, *et al.* PH domain-only protein PHLDA3 is a p53-regulated repressor of Akt. *Cell* **136**, 535-550 (2009).
35. Milne RL, *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* **49**, 1767-1778 (2017).
36. Michailidou K, *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94 (2017).
37. Michailidou K, *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373-380 (2015).
38. Michailidou K, *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-361, 361e351-352 (2013).
39. Rozenblatt-Rosen O, *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236-249 (2020).
40. Day FR, *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat Genet* **49**, 834-841 (2017).

41. Barker DJ, Osmond C, Thornburg KL, Kajantie E, Eriksson JG. A possible link between the pubertal growth of girls and prostate cancer in their sons. *Am J Hum Biol* **24**, 406-410 (2012).
42. Giles GG, *et al.* Early growth, adult body size and prostate cancer risk. *Int J Cancer* **103**, 241-245 (2003).
43. Gann PH, Hennekens CH, Ma J, Longcope C, Stampfer MJ. Prospective study of sex hormone levels and risk of prostate cancer. *J Natl Cancer Inst* **88**, 1118-1126 (1996).
44. Key TJ. Hormones and cancer in humans. *Mutat Res* **333**, 59-67 (1995).
45. Bonilla C, *et al.* Pubertal development and prostate cancer risk: Mendelian randomization study in a population-based cohort. *BMC Med* **14**, 66 (2016).
46. Chan JM, *et al.* Plasma insulin-like growth factor-I and prostate cancer risk: a prospective study. *Science* **279**, 563-566 (1998).
47. Grossmann M, Cheung AS, Zajac JD. Androgens and prostate cancer; pathogenesis and deprivation therapy. *Best Pract Res Clin Endocrinol Metab* **27**, 603-616 (2013).
48. Rowlands MA, Gunnell D, Harris R, Vatten LJ, Holly JM, Martin RM. Circulating insulin-like growth factor peptides and prostate cancer risk: a systematic review and meta-analysis. *Int J Cancer* **124**, 2416-2429 (2009).
49. Salonia A, *et al.* Circulating sex steroids and prostate cancer: introducing the time-dependency theory. *World J Urol* **31**, 267-273 (2013).
50. Hur J, Giovannucci E. Racial differences in prostate cancer: does timing of puberty play a role? *Br J Cancer* **123**, 349-354 (2020).
51. Sandhu J, Davey Smith G, Holly J, Cole TJ, Ben-Shlomo Y. Timing of puberty determines serum insulin-like growth factor-I in late adulthood. *J Clin Endocrinol Metab* **91**, 3150-3157 (2006).
52. Sutcliffe S, Colditz GA. Prostate cancer: is it time to expand the research focus to early-life exposures? *Nat Rev Cancer* **13**, 208-518 (2013).
53. Travis RC, *et al.* A Meta-analysis of Individual Participant Data Reveals an Association between Circulating Levels of IGF-I and Prostate Cancer Risk. *Cancer Res* **76**, 2288-2300 (2016).

54. Axelrad JE, Lichtiger S, Yajnik V. Inflammatory bowel disease and cancer: The role of inflammation, immunosuppression, and cancer treatment. *World J Gastroenterol* **22**, 4794-4801 (2016).
55. Biancone L, *et al.* Cancer Risk in Inflammatory Bowel Disease: A 6-Year Prospective Multicenter Nested Case-Control IG-IBD Study. *Inflamm Bowel Dis* **26**, 450-459 (2020).
56. Burisch J, Munkholm P. The epidemiology of inflammatory bowel disease. *Scand J Gastroenterol* **50**, 942-951 (2015).
57. Hovde O, Hoivik ML, Henriksen M, Solberg IC, Smastuen MC, Moum BA. Malignancies in Patients with Inflammatory Bowel Disease: Results from 20 Years of Follow-up in the IBSEN Study. *J Crohns Colitis* **11**, 571-577 (2017).
58. Yadav S, Singh S, Harmsen WS, Edakkanambeth Varayil J, Tremaine WJ, Loftus EV, Jr. Effect of Medications on Risk of Cancer in Patients With Inflammatory Bowel Diseases: A Population-Based Cohort Study from Olmsted County, Minnesota. *Mayo Clin Proc* **90**, 738-746 (2015).
59. Pellino G, *et al.* Relatives of Crohn's disease patients and breast cancer: an overlooked condition. *Int J Surg* **12 Suppl 1**, S156-158 (2014).
60. Riegler G, *et al.* Increased risk of breast cancer in first-degree relatives of Crohn's disease patients. An IG-IBD study. *Dig Liver Dis* **38**, 18-23 (2006).
61. Cominelli F, Pizarro TT. Interleukin-1 and interleukin-1 receptor antagonist in inflammatory bowel disease. *Aliment Pharmacol Ther* **10 Suppl 2**, 49-53; discussion 54 (1996).
62. Eras N, Daloglu FT, Colak T, Guler M, Akbas E. The Correlation between IL-1beta-C31T Gene Polymorphism and Susceptibility to Breast Cancer. *J Breast Cancer* **22**, 210-218 (2019).
63. Miller LJ, *et al.* Interleukin-1 family expression in human breast cancer: interleukin-1 receptor antagonist. *Cancer Invest* **18**, 293-302 (2000).
64. Vijgen L, *et al.* Interleukin-1 receptor antagonist VNTR-polymorphism in inflammatory bowel disease. *Genes Immun* **3**, 400-406 (2002).
65. Wang J, Shi Y, Wang G, Dong S, Yang D, Zuo X. The association between interleukin-1 polymorphisms and their protein expression in Chinese Han patients with breast cancer. *Mol Genet Genomic Med* **7**, e804 (2019).

66. Lee KM, *et al.* Genetic polymorphisms of interleukin-1 beta (IL-1B) and IL-1 receptor antagonist (IL-1RN) and breast cancer risk in Korean women. *Breast Cancer Res Treat* **96**, 197-202 (2006).
67. Zhou J, Yang R. Identification of key pathways and genes shared between Crohn's disease and breast cancer using bioinformatics analysis. *Oncol Lett* **20**, 119 (2020).
68. Chen M, Yuan C, Xu T. An increase in prostate cancer diagnosis during inflammatory bowel disease: A systematic review and meta-analysis. *Clin Res Hepatol Gastroenterol* **44**, 302-309 (2020).
69. Ge Y, Shi Q, Yao W, Cheng Y, Ma G. The association between inflammatory bowel disease and prostate cancer risk: a meta-analysis. *Prostate Cancer Prostatic Dis* **23**, 53-58 (2020).
70. Meyers TJ, *et al.* Association between inflammatory bowel disease and prostate cancer: A large-scale, prospective, population-based study. *Int J Cancer* **147**, 2735-2742 (2020).
71. Pedersen N, Duricova D, Elkjaer M, Gamborg M, Munkholm P, Jess T. Risk of extra-intestinal cancer in inflammatory bowel disease: meta-analysis of population-based cohort studies. *Am J Gastroenterol* **105**, 1480-1487 (2010).
72. Jess T, Horvath-Puho E, Fallingborg J, Rasmussen HH, Jacobsen BA. Cancer risk in inflammatory bowel disease according to patient phenotype and treatment: a Danish population-based cohort study. *Am J Gastroenterol* **108**, 1869-1876 (2013).
73. Jussila A, Virta LJ, Pukkala E, Farkkila MA. Malignancies in patients with inflammatory bowel disease: a nationwide register study in Finland. *Scand J Gastroenterol* **48**, 1405-1413 (2013).
74. Wilson JC, Furlano RI, Jick SS, Meier CR. A population-based study examining the risk of malignancy in patients diagnosed with inflammatory bowel disease. *J Gastroenterol* **51**, 1050-1062 (2016).
75. Burns JA, *et al.* Inflammatory Bowel Disease and the Risk of Prostate Cancer. *Eur Urol* **75**, 846-852 (2019).
76. Khan N, Vallarino C, Lissos T, Darr U, Luo M. Risk of Malignancy in a Nationwide Cohort of Elderly Inflammatory Bowel Disease Patients. *Drugs Aging* **34**, 859-868 (2017).
77. Haddad A, *et al.* Inflammatory bowel disease and prostate cancer risk: A systematic review. *Arab J Urol* **18**, 207-212 (2020).

78. Rais R, *et al.* FOLH1/GCPII is elevated in IBD patients, and its inhibition ameliorates murine IBD abnormalities. *JCI Insight* **1**, (2016).
79. Yao V, Parwani A, Maier C, Heston WD, Bacich DJ. Moderate expression of prostate-specific membrane antigen, a tissue differentiation antigen and folate hydrolase, facilitates prostate carcinogenesis. *Cancer Res* **68**, 9070-9077 (2008).
80. Zhang T, *et al.* An ileal Crohn's disease gene signature based on whole human genome expression profiles of disease unaffected ileal mucosal biopsies. *PLoS One* **7**, e37139 (2012).
81. Kaittanis C, *et al.* Prostate-specific membrane antigen cleavage of vitamin B9 stimulates oncogenic signaling through metabotropic glutamate receptors. *J Exp Med* **215**, 159-175 (2018).
82. Genkinger JM, *et al.* Measures of body fatness and height in early and mid-to-late adulthood and prostate cancer: risk and mortality in The Pooling Project of Prospective Studies of Diet and Cancer. *Ann Oncol* **31**, 103-114 (2020).
83. Discacciati A, Orsini N, Wolk A. Body mass index and incidence of localized and advanced prostate cancer--a dose-response meta-analysis of prospective studies. *Ann Oncol* **23**, 1665-1671 (2012).
84. Cao Y, Ma J. Body mass index, prostate cancer-specific mortality, and biochemical recurrence: a systematic review and meta-analysis. *Cancer Prev Res (Phila)* **4**, 486-501 (2011).
85. Lewis SJ, *et al.* Associations between an obesity related genetic variant (FTO rs9939609) and prostate cancer risk. *PLoS One* **5**, e13485 (2010).
86. Rodriguez C, *et al.* Body mass index, weight change, and risk of prostate cancer in the Cancer Prevention Study II Nutrition Cohort. *Cancer Epidemiol Biomarkers Prev* **16**, 63-69 (2007).
87. Gong Z, *et al.* Obesity, diabetes, and risk of prostate cancer: results from the prostate cancer prevention trial. *Cancer Epidemiol Biomarkers Prev* **15**, 1977-1983 (2006).
88. Baillargeon J, *et al.* The association of body mass index and prostate-specific antigen in a population-based study. *Cancer* **103**, 1092-1095 (2005).
89. Kristal AR, Chi C, Tangen CM, Goodman PJ, Etzioni R, Thompson IM. Associations of demographic and lifestyle characteristics with prostate-specific antigen (PSA) concentration and rate of PSA increase. *Cancer* **106**, 320-328 (2006).

90. Perez-Cornago A, *et al.* Tall height and obesity are associated with an increased risk of aggressive prostate cancer: results from the EPIC cohort study. *BMC Med* **15**, 115 (2017).
91. Fowke JH, *et al.* The association between body size, prostate volume and prostate-specific antigen. *Prostate Cancer Prostatic Dis* **10**, 137-142 (2007).
92. Davies NM, *et al.* The effects of height and BMI on prostate cancer incidence and mortality: a Mendelian randomization study in 20,848 cases and 20,214 controls from the PRACTICAL consortium. *Cancer Causes Control* **26**, 1603-1616 (2015).
93. Richardson TG, Sanderson E, Elsworth B, Tilling K, Davey Smith G. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. *BMJ* **369**, m1203 (2020).
94. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756 (2018).
95. McCarthy S, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016).
96. Buniello A, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
97. Fritsche LG, *et al.* Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am J Hum Genet* **107**, 815-836 (2020).
98. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

749 **ACKNOWLEDGMENT**

750 This work was supported by grants from the U.S. National Institutes of Health (U01CA194393
751 and R01CA227237). The results published here are in whole or part based upon data generated
752 by the TCGA Research Network: <https://www.cancer.gov/tcga>.

753 **AUTHOR CONTRIBUTIONS**

754 P.K., A.G., and Y.J.H. designed the study. P.K. and Y.L. developed the analysis plan. L.B.A.
755 prepared the somatic data of TCGA. A.G. prepared the germline data of TCGA. Y.L. performed
756 the analyses. P.K. and Y.L. wrote the manuscript. A.G., Y.J.H., and L.B.A. reviewed the
757 manuscript.

758 **COMPETING INTERESTS**

759 The authors declare no competing interests.

760 **MATERIALS & CORRESPONDENCE**

761 Correspondence and requests for materials should be addressed to P.K. (email:
762 pkraft@hsph.harvard.edu)

TABLES

Table 1. Significant associations between tumor somatic mutation counts and germline PRS.

Cancer type	Somatic mutation count	Germline PRS ^a	Direction of association ^b	<i>P</i> -value ^c (without tumor stage)	<i>P</i> -value ^d (with tumor stage)
PRAD	SBS1	Age at menarche	-	1.21×10^{-8}	.
PRAD	SBS1	IBD	+	1.94×10^{-5}	.
PRAD	SBS1	CD	+	3.41×10^{-7}	.
PRAD	SBS1	UC	+	2.01×10^{-5}	.
PRAD	SBS1	GBM	-	3.29×10^{-8}	.
PRAD	SBS1	HNSC	-	8.39×10^{-6}	.
PRAD	SBS1	BMI	+	4.95×10^{-9}	.
BRCA	APOBEC-related ^e	BRCA	-	1.14×10^{-5}	7.61×10^{-3}
BRCA	APOBEC-related ^e	BRCA ER+	-	8.64×10^{-6}	4.39×10^{-3}
BRCA	APOBEC-related ^e	IBD	+	6.47×10^{-7}	9.77×10^{-5}
BRCA	APOBEC-related ^e	UC	+	9.27×10^{-6}	7.28×10^{-4}
BRCA	SBS1	CD	+	3.06×10^{-5}	6.23×10^{-5}
BRCA	TSMC	IBD	+	3.13×10^{-5}	5.78×10^{-4}

BRCA ER-	SBS1	IBD	+	2.86×10^{-5}	1.31×10^{-4}
BRCA ER-	TSMC	UC	+	9.72×10^{-6}	3.42×10^{-4}
UCEC	SBS40	IBD	-	8.53×10^{-6}	.
UCEC	SBS40	CD	-	1.34×10^{-6}	.

^a PRS is calculated from the germline genetic data of the same TCGA patient as the tumor sample.

^b Direction of the association between somatic mutation count and PRS from zero-inflated negative binomial model, negative binomial model, or linear model. The directions are the same for model adjusting for tumor stage and model not adjusting for tumor stage for all breast cancer associations.

^c *P*-value associated with PRS without adjusting for tumor stage. *P*-values are obtained from likelihood ratio test of model with PRS and model without PRS. For zero-inflated negative binomial model, the results are from testing the count and logistic model jointly.

^d *P*-value associated with PRS adjusting for tumor stage. *P*-values are obtained from likelihood ratio test of model with PRS and model without PRS. For zero-inflated negative binomial model, the results are from testing the count and logistic model jointly. Tumor stage for PRAD and UCEC cases are not available from the GDC data portal <https://portal.gdc.cancer.gov/>.

^e APOBEC-related signatures count is the sum of SBS2 and SBS13 mutation counts, both signatures are attributed to the enzymatic activity of the APOBEC family of cytidine deaminases.

Abbreviations: PRS = polygenic risk score; TSMC = total somatic mutation counts; SBS = single-base substitution; BRCA = breast invasive carcinoma; BRCA ER- = breast invasive carcinoma without estrogen receptors; GBM = glioblastoma multiforme; HNSC = head and neck squamous cell carcinoma; PRAD = prostate adenocarcinoma; UCEC = uterine corpus endometrial carcinoma; BMI = body mass index; IBD = inflammatory bowel disease; CD = Crohn's disease; UC = ulcerative colitis

FIGURES

Fig 1. Mutation counts of SBS signatures and TSMC across 12 TCGA cancer types and two subtypes of breast cancer. Each dot represents a tumor sample. The median of $\log_{10}(\text{mutation count}+1)$ for each cancer type and SBS signature (or TSMC) is represented by both the color of points and the black short lines in each panel. Sample size for each cancer type is shown on the

top.

Abbreviations: TSMC = total somatic mutation counts; SBS = single-base substitution; BLCA = bladder urothelial carcinoma; BRCA = breast invasive carcinoma; BRCA ER+ = breast invasive carcinoma with estrogen receptors; BRCA ER- = breast invasive carcinoma without estrogen receptors; COAD = colon adenocarcinoma; GBM = glioblastoma multiforme; HNSC = head and neck squamous cell carcinoma; KIRC = kidney renal clear cell carcinoma; LGG = brain lower grade glioma; LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma; OV = ovarian serous cystadenocarcinoma; PRAD = prostate adenocarcinoma; UCEC = uterine corpus endometrial carcinoma.

Fig. 2. Correlations between somatic mutation counts and age at cancer diagnosis across cancers. Only the correlations with age at cancer diagnosis for those cancer-signature pairs included in the analyses are shown in this figure. Number in each cell and the cell color represent the Spearman correlation (ρ) between mutation counts of a SBS signature or TSMC (y-axis) and age at diagnosis in a cancer type (x-axis). Corrections passed the Bonferroni threshold ($p < 0.05/69 = 0.00725$) are marked with triple asterisk (***), correlations with $p < 0.01$ are marked with double asterisk (**), and correlations with $p < 0.05$ are marked with single asterisk (*).

SBS1 and CD PRS

SBS1 and bladder cancer PRS

APOBEC-related signatures and CD PRS

Fig. 3. The associations between somatic mutation counts of SBS1 and CD PRS, SBS1 and bladder cancer PRS, and APOBEC-related signatures and CD PRS across cancers. The fixed-effect and Stouffer's p-values for the association between SBS1 and CD PRS are $p = 6.46 \times 10^{-6}$ and $p = 1.25 \times 10^{-5}$; for the association between SBS1 and the bladder cancer PRS, they are $p = 8.30 \times 10^{-5}$ and $p = 4.01 \times 10^{-4}$; and for the association between APOBEC-related signatures and CD PRS, they are 3.41×10^{-4} and $p = 3.92 \times 10^{-4}$). There are significant heterogeneities in the effect sizes for all three associations across cancers ($p < 0.01$). The effect sizes and 95% CI for PRS are plotted using grey squares and black horizontal lines. Size of the grey squares represents the weight in fixed-effect model for each cancer type. Red dashed line and red diamond represent the results from fixed-effect model.

Fig. 4. Hypothetical relationships between germline PRS, risk factor, mutational signature, and diagnosis of cancer. **a** Cancer PRS is associated with mutational signature due to collider bias. Cancer diagnosis (D) may or may not have an effect on somatic mutations of certain mutational signature (M). The cancer risk factor (X) is independent of cancer PRS (G) in the general population and is also associated with mutational signatures (M). Conditioning on diagnosis (D , i.e., studying cancer cases only) would induce collider bias on the relationship between cancer PRS (G) and mutational signature (M). **b** Three possible relationships between cancer or non-cancer PRS, mutational signature, and cancer diagnosis assuming no reverse causation. **(i)**

Indirect effect: PRS (G) has an indirect effect on tumor development and diagnosis (D) through inducing somatic mutations of certain mutational signature (M); **(ii) Non-carcinogenic effect:** PRS (G) has an effect on inducing somatic mutations of certain mutational signature (M) but neither PRS (G) nor somatic mutations (M) has an effect on tumor development and diagnosis (D); **(c) Direct (and indirect) effect:** PRS (G) has a direct effect on tumor development and diagnosis (D) that not through the effect of somatic mutations (M) and may or may not have an indirect effect through somatic mutations (M). In this case, conditioning on diagnosis (D , i.e., studying cancer cases only) would induce collider bias on the relationship between PRS (G) and mutational signature (M).