Attribution of Cancer Origins

to Endogenous, Exogenous, and Actionable Mutational Processes

Vincent L. Cannataro[1] (https://orcid.org/0000-0002-6364-7747),

Jeffrey D. Mandell[2] (https://orcid.org/0000-0002-3839-2543),

Jeffrey P. Townsend[2,3,4†] (https://orcid.org/0000-0002-9890-3907)

[1]Department of Biology, Emmanuel College, Boston, MA,

[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT

[3]Department of Biostatistics, Yale School of Public Health, New Haven, CT,

and

[4]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT

[†]To whom correspondence should be addressed (Jeffrey.Townsend@Yale.edu).

Keywords: cancer, tumor, single-nucleotide variants, mutational signatures, effect size, somatic mutation, selection, evolution, prevention, public health, molecular epidemiology

SUMMARY

Tumor genomes are suffused with mutations, each caused by specific mutational processes that can leave tell-tale signatures revealing their aetiology. However, knowledge of these underlying mutational processes acting on our DNA does not on its own convey the causation of cancer: many mutations are "passenger" mutations that do not affect progression. One must determine which mutations are causative, which are not, and to what degree. Here we directly tie the cancer effect size—a metric of the proliferation and survival benefit conferred by molecular variants—to the mutational processes that generated molecular variation within individual tumors, quantifying the contribution of each mutagenic process to the formation of each whole-exome sequenced tumor. Furthermore, we extend this result at the individual level to demonstrate the specific mutational processes that are most responsible for tumors of 23 major cancer types. Melanomas and lung cancers are attributable to the actionable, preventable, exogenous mutational processes of ultraviolet light and tobacco exposure, whereas gliomas and prostate adenocarcinoma tumors are largely attributable to endogenous processes associated with aging. Other potentially actionable mutational processes, such as mutations associated with pathogen exposure and APOBEC activity, account for a large proportion of the cancer effect within head and neck, bladder, cervical, and breast cancers. These molecular attributions complement and expand longstanding epidemiological approaches that inform cancer prevention by revealing the burden of cancer caused by either endogenous or exogenous, non-actionable or actionable processes.

Introduction

In the past half-century, our understanding of the origins of cancers has progressed from a deep mystery to a widespread acceptance that cancers are the outcome of an evolutionary process driven by mutation, consequent genetic variation, and selection on that genetic variation (*1–3*). Epidemiological studies have established both an association with age (*4*) and causation by exposure to carcinogens (*5*), demonstrating that endogenous processes and exogenous mutagens can increase the rate of mutations (*6*), create somatic genetic variation (*7*), and increase the rate of cancer (*8*, *9*). In recent years, large-scale analyses have been able to recover characteristic tissue-specific signatures of these underlying mutagenic processes in the variants revealed by whole-exome and whole-genome cancer tumor sequencing (*10–12*). However, the specific cancer driver architecture within each kind of cancer tissue has also been demonstrated to be both predictable (*13*) and, importantly, circumscribed (*14*). Therefore, the causation of cancer by each mutational processes is not determined solely by their effect on mutation rate nor upon the amount of somatic genetic variation they induce, but critically depends upon the degree to which the specific mutations they supply provide selective advantages to clonal lineages within tissues that give rise to cancer.

To evaluate selective advantages requires knowledge of mutation bias, for which characteristic patterns have long been attributed to specific tissues (*15–18*). These patterns can be validated within model organisms under laboratory conditions and attributed to specific mutagenic sources (*19*); numerous algorithms have been developed to deconvolve the total substitution load to its constituent mutational signatures (*20*). Application of these algorithms to whole-exome or whole-genome data recapitulates underlying trinucleotide mutation rates without bias from natural selection because the vast majority of mutations are accumulating neutrally (*21*, *22*). Nevertheless, Poulos et al (*23*) demonstrated that known major driver mutations are statistically associated with specific mutational signatures. Therefore, specific mutagenic processes in different tissues are driving tumorigenesis via mutations in genes that confer a survival and proliferative advantage to somatic cells. Estimation of the effects of each mutagenic process on the development of cancer requires quantification of the effects of each SNV toward tumorigenesis.

Quantification of the cancer effect of mutations requires estimation of their relative impact on cancer lineage survival and replication, an estimation that critically depends on an understanding of the baseline rate of mutation in the absence of natural selection (*22*). Ostrow et al. (*24*) performed a comprehensive analysis of ratios of non-synonymous change to synonymous change to quantify genomic natural selection in the somatic evolution of cancer. This approach has been applied in the field of evolutionary biology for decades and has recently been adapted to the nuances of cancer evolution in several meaningful ways (*25*, *26*), such as taking tissue-specific trinucleotide mutational patterns into account (c.f. *27*). Martincorena et al. (*28*)

performed an analysis using trinucleotide substitution rates and covariate-informed gene-level mutation rates to quantify gene-wide selection conferring enhanced proliferation and survival of cancer cell lineages. Temko et al. (*29*) deconvolved the underlying mutational signatures in tumor sets, associated signatures and drivers, and quantified the relative intragenic selection of the SNVs in a selection of high-burden driver genes. Cannataro et al. (*30*) quantified the site-specific selective effect on each SNV during primary tumor development by determining the constituent mutational signatures driving mutation load in each tumor, coupling these rates with covariate-informed gene-level mutation rates, and quantifying their contribution to cancer cell lineage survival and reproduction in comparison to the convolved baseline mutation rate. These cancer drivers—and their relative effect—may be related back to the mechanisms driving genomic variation, i.e., the processes behind the detected mutational signatures.

Epidemiological studies have correlated environmental exposures to specific cancer incidences for decades (*31*, *32*). Recently, cancer incidence has also been correlated with tissue-specific stem cell division numbers (*33*, *34*), which has been interpreted as evidence that cancers are mainly driven by endogenous, i.e., aging or "bad luck", effects. Other analyses dispute this conclusion, pointing out that it is confounded by the sensitivity of rapidly dividing tissues to exogenous mutational sources (*35*, *36*), and by the exclusion of cancer types with known environmental causes (*37*). To determine the relative contributions of endogenous and exogenous processes on cancer phenotypes, tumor sequence data can be used to parameterize the magnitude of age-associated, exogenous and actionable mutational processes and their cancer effects of age-associated and actionable mutations on tumorigenesis. Such analyses of the evolutionary dynamics driving tumorigenesis back to the sources of the heterogeneity fueling cancer evolution are essential to the advancement of our understanding of oncogenesis and cancer prevention.

Here we analyze the signatures of mutational processes in diverse cancer types. We quantify the cancer effect size of consequent single-nucleotide variants. By computing the relative contribution of diverse mutational processes to cancer effect, we determine the degree of cancer effect attributable to each mutagenic process. We quantify the contribution of each mutagenic process to cancer in individual patient tumors, and their relative contribution across tumors within sampled cancer types. We identify cancer types where the discrepancy between mutagenic input and cancer effect is largest, and smallest, and analyze which mutagenic processes are most proportionally discrepant with their cancer effect within each cancer type. This analysis enables comparison of the proportions of cancer effect attributable to age-associated processes to the proportions of cancer effect attributable to putatively preventable mutagenic processes such ultraviolet light exposure, tobacco smoking or chewing, and APOBEC mutagenesis, addressing a longstanding controversy regarding the role of endogenous luck and exogenous exposure to tumorigenesis—and moreover, informing the benefits of prevention of mutation in the prevention of cancer.

Methods

To attribute the increased cellular reproduction and survival conferred by single nucleotide variants responsible for cancer growth to their underlying mutational sources we determined the sources of mutation within individual tumors, calculated the effect size of each single nucleotide substitution among tumors in each tumor type, and evaluated the likelihood that each of these substitutions was the product of each mutational source within each tumor. Thus, single-nucleotide substitutions responsible for the largest influence on cellular division and survival, and hence the cancer phenotype, may be attributed to the root sources of molecular variation within each somatic tissue. We analyzed the pan-cancer whole-exome tumor sequencing dataset curated in (*30*), except all Yale-Gilead tumors that might have been treated with chemotherapies were removed (removed tumors in **Table S1**). Scripts used to perform these analyses are available online at https://github.com/Townsend-Lab-Yale/cancer_causes_and_effects.

Attributing Sources of Mutation within Tumors

To attribute observed sets of substitutions in tumors to the underlying sources of mutations, we used the `R` package `deconstructSigs` (*38*) to extract version 3 COSMIC mutational signatures from each tumor's set of non-recurrent substitutions. (We excluded recurrent variants because they are much more likely to be under selection in the cancer cell population; non-recurrent mutations more accurately reflect mutational influx.) As some COSMIC signatures share similar mutational profiles, we limited the number of signatures detectable in each tumor type to those signatures detected at any prevalence in tumors of that type previously (*10*) to minimize signature bleeding. We also applied the recommended minimum threshold for the number of substitutions necessary to attribute to a signature associated with increased mutagenesis. For example, signatures attributable to defective DNA mismatch repair were only allowed in tumors with over 200 substitutions (*10*). Some tumors analyzed exhibited fewer than 50 substitutions (**Supplemental Fig. S1**)—a threshold below which precise deconvolution of mutational signatures becomes problematic (*38*). For these tumors, we mixed the `deconstructSigs` estimates of the signature weights for the specific tumor with the average signature weights for the tumors with 50 or more substitutions of the same tumor type, weighting the former in proportion to the number of variants in the tumor out of 50.

As some COSMIC signatures have been attributed to artifactual processes such as sample handling and sequencing, we focus on the tumor-type-specific subset of signatures B that represent biologically relevant mutational processes (*10*). The fitted weights of signatures in B reflect the relative rates that their underlying mutational processes contribute mutations. To determine the tumor-specific relative weight w_i of a biological signature $i \in B$, we divided its

fitted weight $\widehat{w}_i$ by the sum of the fitted weights of all biologically associated signatures; i.e.,

$$w_i = \frac{\widehat{w}_i}{\sum_{b \in B} \widehat{w}_b} . \tag{1}$$

Calculating the mutational source variant weight

After calculating the tumor-specific relative weights of mutational sources (**Fig. 1A**) as described in **Eq. 1** (**Fig 1B**), we used the trinucleotide-context-specific relative mutation rates defined for mutational signatures to calculate the probabilities that each variant in each tumor derives from the mutational processes underlying each biological mutational signature. Let $\boldsymbol{\Psi}$ be a matrix of trinucleotide-context-specific relative mutation rates for the signatures in B, with $\psi_{i,j}$ being the rate for signature i of trinucleotide-context-specific mutation j. Retrospectively, the probability that a single-nucleotide variant constituting a trinucleotide-context-specific mutation j derived from the process underlying mutational signature i in tumor n is

$$p_{i,j,n} = \frac{w_{i,n}\psi_{i,j}}{\sum_{b \in B} w_{b,n}\psi_{b,j}} , \tag{2}$$

where each $w_{k,n}$ is the relative weight of signature k in tumor n (**Eq. 1**).

For a given variant, we can scale these probabilities by the impact of the variant on the survival and proliferation of cancer cells—that is, the variant's cancer effect size—to quantify the relative contributions of each mutational source to the cancer.

Calculating the Effect Sizes of Variant Substitutions in Tumors

To attribute a cancer effect to each substitution, we used the R package cancereffectsizeR, version 0.1.1.9010 (*30*). As described in our previous work, the package's underlying model assumes that substitutions fix in accord with a Poisson distribution at the rate that mutations arise (mutation rate μ) multiplied by their cancer effect size γ. The term γ is $\frac{\lambda}{\mu} = N \times u(s)$, where λ is the rate of substitutions, N is the effective population size of cancer cells, and $u(s)$ is the probability of fixation of a new mutation as a function of the selection coefficient s, from population genetic theory. Thus, for every variant, we calculated the effect size that maximized the likelihood function

$$\mathcal{L}(\gamma|\mu_1, \ldots, \mu_N; N_j, N_k) = \prod_{j=1}^{N_j}(1 - \mathrm{e}^{-\mu_j\gamma}) \times \prod_{k=1}^{N_k} \mathrm{e}^{-\mu_k\gamma} \tag{3}$$

for the N tumors, N_j of which exhibited at least one substitution event of that variant in tumor j, and N_k of which exhibited no such variant in tumor k. Each tumor-specific mutation rate $\mu_1, \ldots, \mu_N$ was calculated by extracting the trinucleotide mutation rate of each variant from the tumor-specific mutational signature weights (**Eq. 1**) and convolving it with the gene-specific mutation rate as in Cannataro et al. (*30*).

Calculating mutational source effect weight

The relative contribution to cancer effect of mutational process *b* in tumor *n*

$$\alpha_{b,n} = \frac{\gamma_i p_{i,n,b}}{\sum_{i=1}^{N} \gamma_i p_{i,n,b}} \tag{4}$$

scales the effect size γ_i of variant *i* by the probability it was caused by the process underlying signature *b*, within each tumor, relative to all contributions of scaled cancer effect in that tumor.

Lastly, we sum $\alpha_{b,n}$ across all tumors in a given tumor type to get the population-level burden of scaled cancer effect size contributed by signature *b*.

Results

Proportionate Contributions of Mutational Processes to Cancer Effect Can Be Calculated

To determine the sources of mutagenesis occurring in tumor samples, we deconvolved the mutational burden of each tumor into the most likely distribution of attributed single-nucleotide variant mutational signatures (*39*, *40*). Applied to a lung squamous cell carcinoma (LUSC) tumor sample from a single patient (MDA-1229-T), this deconvolution yielded three trinucleotide mutational signatures (age-associated #5, tobacco smoking #4, and unknown aetiology #8; **Fig. 1A**), each contributing to the flux of single nucleotide variants in the tumor at a calculated weight (**Fig. 1B**). The trinucleotide signature weight or combination of trinucleotide signature weights contributing to a specific variant (**Fig. 1A**) times the proportion of mutational causation attributable to each corresponding cancer effect (**Fig. 1B**) provides the probability each source contributed to each variant in this tumor (**Fig. 1C**). In this instance, age-associated signature #5 is the most likely contributor to NFE2L2 R34P and TP53 R282W, whereas tobacco smoking is the most likely source of mutations causing OR13G1 L79L. However, only a few of the mutations that occur in somatic tissue are thought to be selected for their effects on growth or survival, and therefore causative of cancer, and the level of causation is presumably quantitative—i.e., the mutations in a type of cancer that drive cancer are responsible to different degrees for the manifestation of a cancer phenotype (*30*). In this case, NFE2L2 R34P has a higher cancer effect size than TP53 R282W, and the odorant receptor mutation OR13G1 has negligible to no effect. The product of the probability that each mutational source contributed to each variant in this tumor and the effect of the specific variant (**Fig. 1D**) quantifies the probability-weighted cancer effect for each variant by each source (**Fig. 1E**). Summing the probability-weighted cancer effect for each source across variants yields the proportion of cancer effect attributable to each source of mutations (**Fig. 1F**). Age-associated mutational signature #5 contributed the highest weight in MDA-1229-T, and led to the largest estimated effect through its high probability of being causative of both the NFE2L2 R34P and TP53 R282W mutations. Via deconvolution of the mutational signatures responsible for recurrent variants in cancer and

calculation of the cancer effect sizes of the nucleotide substitutions driving cancer evolution, we have calculated which mutagenic sources fueling nucleotide variation can be attributed as proportionally causative of individual tumors in patients.

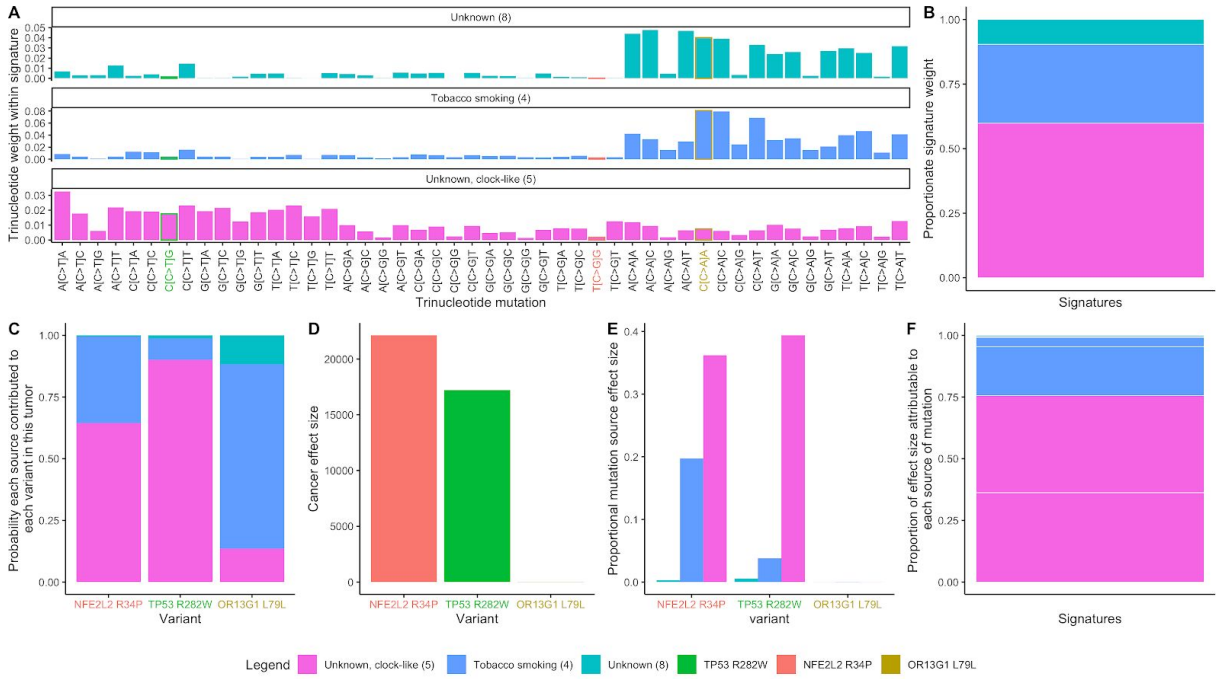


Figure 1: Calculating the cancer effects by mutational source. The respective products of (**A**) the trinucleotide weights within each signature (e.g. a relatively infrequent CCG→CTG mutation leads to TP53 R282W, green; a relatively infrequent TCG→TGG mutation leads to NFE2l2 R34P, orange; and a relatively frequent CCA→CAA mutation leads to OR13G1 L79L, tan) and (**B**) the proportions of observed mutations in a tumor caused by each signature (aquamarine: damage by reactive oxygen species; blue: unknown; pink: tobacco smoking) can be normalized to yield (**C**) the probability each source contributed to each variant in this tumor. Each of these probabilities serves as a weight to multiply by (**D**) the cancer effect size of each variant, to yield (**E**) the probability-weighted portion of effect size for each variant attributable to each source of mutations, stacked to compose (**F**) the proportion of cancer causation attributable to each source of mutations. This example illustrates the calculation on a lung squamous-cell carcinoma, MDA-1229-T.

Mutagenic Input and Cancer Effect from each Source Can Differ Substantially within Tumors

The match between the proportional input to total mutations by each mutagenic source (**Fig. 1B**) and the proportional cancer effect arising from each mutagenic source (**Fig. 1F**) varies in each patient's tumor (**Fig. 2**). Quantifying the match by the Jensen-Shannon Divergence (JSD) between proportional mutational input and proportion of cancer causation, we found the tumor type with the lowest median divergence to be ovarian serous cystadenocarcinoma (OV, **Fig. 2A**). The mutational input to the OV tumor in **Fig. 2A** with the lowest JSD (TCGA-24-1103) was

entirely attributed to the BRCA-1- and BRCA-2-associated signature (#3); thus, all cancer effects were attributable to this single source of mutation. Indeed, tumor sample TCGA-24-1103 has a somatic BRCA2 L1638E mutation. Examining a tumor at the second quartile of the distribution of JSD (TCGA-57-1992; **Fig. 2A**), there is a slight mismatch between the mutational input and the contribution to cancer causation—with the BRCA-1- and BRCA-2-associated signature (#3) exhibiting slightly less cancer effect than signature weight (TCGA-57-1992 has a somatic BRCA2 L1195R mutation). Three additional tumors drawn at the median JSD, the fourth quintile, and the highest JSD demonstrate increasing degrees of mismatch between the mutational input and the contribution to cancer causation. These mismatches are even more frequent in 22 other tumor types analyzed (rectal adenocarcinoma, at approximately the second quartile of median JSD across tumor types, **Fig. 2B**; human papillomavirus virus (HPV)-negative head and neck squamous-cell carcinoma, at approximately the third quartile of median JSD across tumor types, **Fig. 2C**; and low-grade glioma, exhibiting the greatest median mismatch across tumor types, **Fig. 2D**; **Supplemental Fig. 2**).

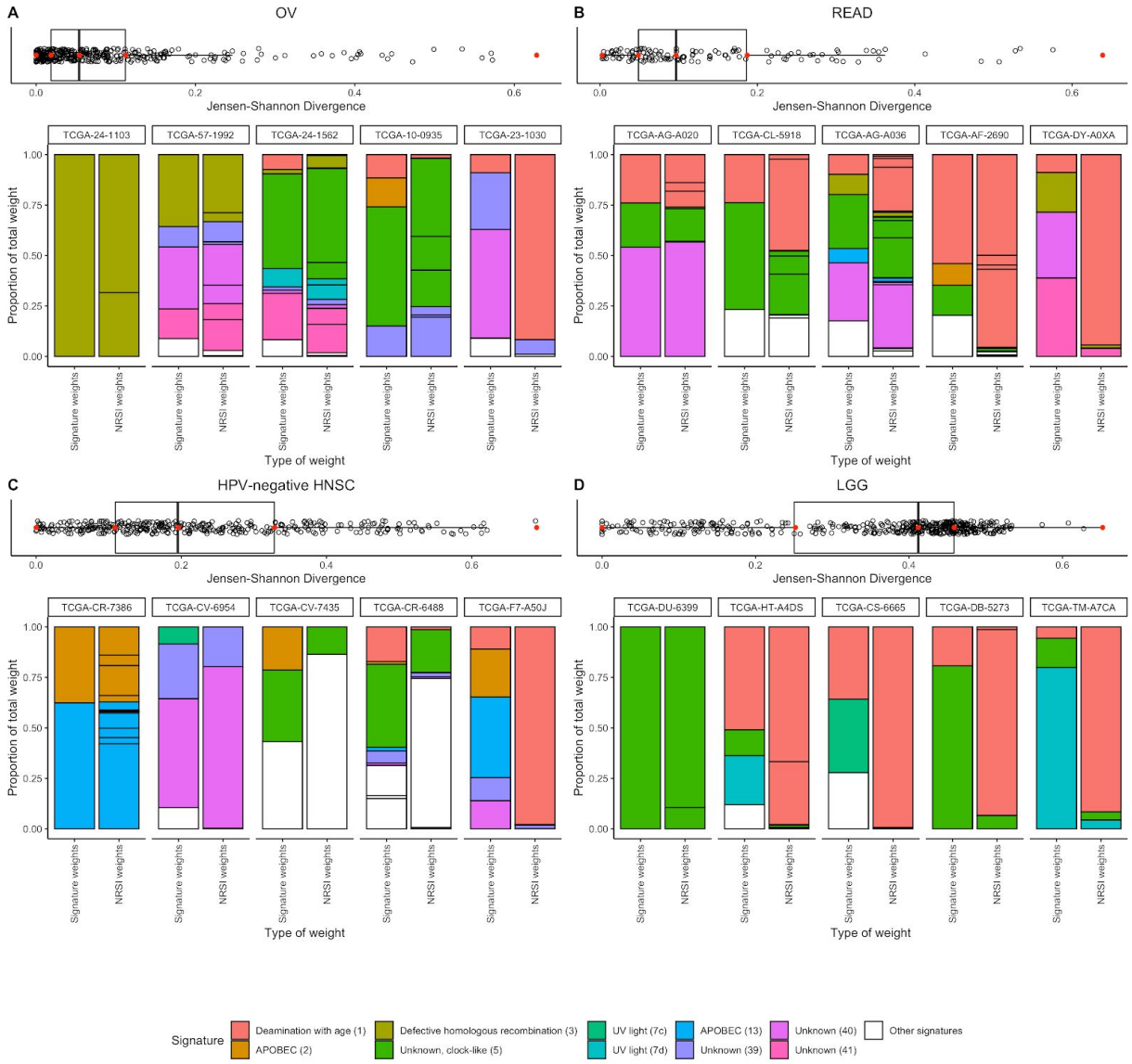


Figure 2. Box plots of the Jensen-Shannon Divergence (JSD) between the proportional input of each mutagenic source and the proportion that each signature contributed to the total cancer effect in each tumor, accompanied by the proportions of observed mutations in a tumor caused by each signature and the proportions of cancer-causation attributable to each source of mutations, for the five tumors (red dots) bounding the quartiles of the JSD among tumors, for four cancer types that bound the tertiles of median JSD among cancer types: (**A**) the cancer type with the least median divergence between the proportional input of each mutagenic source and the proportion that each signature contributed to the total cancer effect, ovarian serous cystadenocarcinoma, (**B**) HPV-negative head-and-neck squamous-cell carcinoma, (**C**) prostate adenocarcinoma, and (**D**) the cancer type with the greatest median divergence between the proportional input of each mutagenic source and the proportion that each signature contributed to the total cancer effect, lower-grade glioma.

Relative Mutagenic Input and Relative Cancer Effect are Specific to each Tumor Type

The mismatches between the proportional input to total mutations by each mutagenic source (**Fig. 1B**) and the proportional cancer effect arising from each mutagenic source (**Fig. 1F**) exist not only at the level of individual tumors, but also at the level of tumor types—where they indicate which mutational sources make an outsized contribution to the causation of cancer compared to their production of mutations, and *vice versa*. Many tumor-type mutational-signature pairs exhibit statistically significant differences between the proportional input to total mutations by each mutagenic source and the proportional cancer effect arising from each mutagenic source (continuity-corrected Wilcoxon two-sided rank-sum tests, $P < 0.05$; **Fig. 3A**). For example, APOBEC-related signatures 2 and 13 exhibit larger mutation weight than cancer effect across many cancer types, as do 17a, 21, 22, 26, 28, and 37. In contrast, the aging-associated signature 1 exhibits larger cancer effect than mutation weight across many cancer types, as does polymerase-epsilon signature 10b. In lower-grade glioma, the age-associated signature 5 constitutes much more of the mutation weight than its cancer effect (66% compared to 23%). This difference between the two age-associated signatures is largely attributable to the high effect size of IDH1 variants, which occur predominantly as a consequence of ACG→ATG mutations that are frequent in signature 1 and rare in signature 5. A similar contrast can be seen in thyroid adenocarcinoma, wherein the APOBEC-related signature 2 exhibits high mutation weight and virtually zero cancer effect (30% compared to 2%), and wherein the aging signature 5 exhibits much more cancer effect than mutation weight (68% compared to 30%). This contrast comes about because thyroid adenocarcinoma is often driven by BRAF V600E mutations that convey enormous cancer effects, and BRAF V600E mutations come about frequently as a consequence of GTG→GAG trinucleotide mutations that are found at low frequency within the aging signature 5, but are found at extremely low frequency within APOBEC signature 2.

Preventable Mutational Processes Contribute Substantially to Causation of Skin, Head-and-Neck, Cervical, and Lung Cancer

Among the non-age-related etiologies are a number of mutational processes that are putatively "preventable"—in that they can be mitigated by individual behaviors or interventions (**Fig. 3B–C**). Skin cancer, lung cancer, HPV-positive head and neck cancer, and cervical cancer are notable for the dominant role of putatively preventable processes underlying both raw SNV mutation weight (**Fig. 3A**) and cancer effect (**Fig. 3B**). Lower-grade glioma, glioblastoma, and

prostate adenocarcinoma are notable for the lack of putatively preventable processes underlying both raw mutation weight and cancer effect. However, the mutation weights and cancer effects are not the same. For example, 57% of the mutation weight of thyroid adenocarcinoma is associated with APOBEC processes that might be preventable by avoiding viral infections. However, these mutations contribute only 2.9% of the cancer effect ($P < 0.001$, Wilcoxon rank-sum test), so that preventing APOBEC-associated mutation would likely do little to prevent cancer. A contrasting case is the lung cancers: the net cancer effects of the SNVs attributable to tobacco chewing (3% in LUAD and 4% in LUSC) and tobacco smoking (44% in LUAD and 24% in LUSC) are larger than the mutation weights of these sources of mutagenesis (2% and 2% for tobacco chewing and 35% and 19% for smoking in LUAD and LUSC, respectively; $P < 0.001$ for each tobacco signature—4 and 29—for both lung cancer types; Wilcoxon rank-sum test).

Age-associated Mutational Processes Contribute Substantially to Causation of Glioma, Prostate, Thyroid, Pancreatic, and Colorectal cancer

Because each mutagenic process is linked to a trinucleotide variant signature that has been identified as clock-like (ubiquitous and age-associated) or non-clocklike (*10*, not associated with age; *41*), the proportion of total mutations attributable to clock-like processes and non-clocklike processes in each cancer type can be quantified (**Fig. 3B**). Among tissues, the cancer types with the greatest proportion of total mutations contributed by non-clocklike processes are melanoma (primary and metastatic), cervical squamous cell carcinoma and endocervical adenocarcinoma, and head and neck cancers. Lower grade glioma exhibits the greatest proportion of total mutations contributed by age-associated, "clocklike" processes. Moreover, with regard to the explanation of tumorigenesis and cancer incidence, the cancer effect attributable to age-associated processes and non-age-associated processes in each cancer type can be quantified (**Fig. 3C**). Among tumor tissues, those with the greatest proportion of cancer effect contributed by age-associated processes are gliomas (LGG, GBM) and prostate adenocarcinomas, consistent with the strong association of the incidence of these cancers with age (*42*, *43*), as well as thyroid and pancreatic cancers. Primary and metastatic melanoma, lung adenocarcinoma, and HPV-positive head and neck squamous-cell carcinoma exhibit the greatest proportion of cancer effect contributed by non-age-associated processes, consistent with the strong association of these cancers with exogenous factors (UV exposure, smoking, and HPV infection).

Our analysis attributes an amount of cancer causation to such endogenous and inactionable processes that varies widely among cancer types. Cancer types varied in the degree to which their causation was associated with COSMIC signature #1, which correlates with stem cell division in different tissues and represents the processes associated with the mitotic clock (**Fig. 3C**; c.f. *33*, *34*), ranging from extremely small contributions in THCA, LUSC, SKCM, KIRC, LUAD, and LIHC, to 76% of the cancer effect in LGG. Even combining both the replication-associated and non-replication-associated aging signatures, causation attributable to

all aging-associated processes ranged from 10% in SKCM to 98% in LGG; around half of the cancer types studied were less than half-caused by age-associated processes. Age has not been associated across cancer types with any of the signatures that have unknown etiology; OV, KIRC, LIHC, BRCA (ER-), and all have >50% of their cancer effect caused by currently unknown or unattributed mutational processes (**Fig. 3C**).

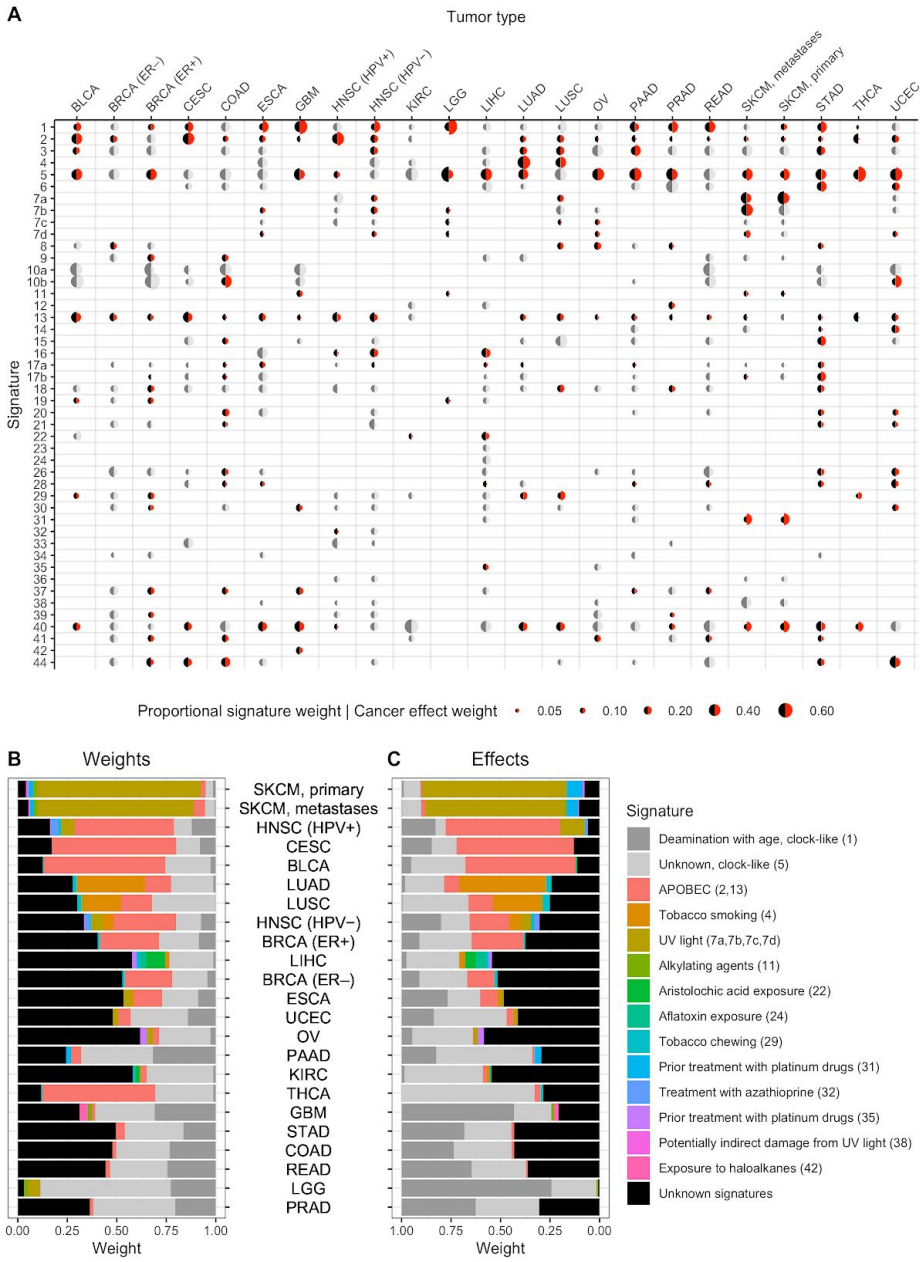


Figure 3. Mutation weights and cancer effects for unknown and etiology-associated mutation signatures. (**A**) Relative contributions of processes attributed to each mutational signature across 23 cancer types to total SNVs (black half-circle) and cancer effects (red half-circle). For each cancer

type, some mutagenic processes contribute lesser proportions of cancer effect than the proportion of SNVs that they cause (e.g. SBS1 in PAAD; 5 in LGG; 7a and 7b in SKCM; 13 and 40 in THCA), and some mutagenic processes contribute greater proportions of cancer effect than the proportion of SNVs that they cause (e.g. SBS1 in ESCA, HNSC, LGA, PRAD, and STAD; 2 in HPV^+ HNSC; 5 and 40 in THCA). (**B**) Relative contributions of age-associated (grey), actionable (colors), and unknown processes (black) across 23 cancer types to total SNVs. (**C**) Relative cancer effects of age-associated, actionable, and unknown processes.

Discussion

Here we have shown that the impact on carcinogenesis of mutagenic processes associated with single-nucleotide variant signatures can be quantified. This quantification is distinct from the number or proportion of mutations that can be attributed to a process, because it accounts for the extent to which each mutation contributes to the cancer phenotype—increased replicative and survival advantage in each tissue and cancer type—via single-nucleotide variants. We have shown how to use the proportions of observed mutations in a tumor caused by each signature to calculate the probability that each mutational source contributed to each variant in this tumor. Each of these probabilities serves to weight the cancer effect size of each variant, yielding the probability-weighted portion of effect size for each variant attributable to each source of mutations and thus the proportion of cancer-causation attributable to each source of mutations. In turn, the quantification of cancer-causation within each tumor, characterized across a population of patients, provides a reductionist molecular approach toward quantifying the degree to which a process can be held responsible for carcinogenesis in a cancer type that is wholly distinct from traditional epidemiological studies.

Our analysis of the cancer effects of single-nucleotide mutations and associated signatures has been enabled by quantitative estimates of their intrinsic mutation rates (*44–46*). Germline variants, copy-number variation, epigenetic alterations, and changes to the aging tissue microenvironment also contribute to the cancer phenotype (*47–52*). Incorporation of signatures associated with these kinds of alterations (*53*) and of attributions of each signature to relevant sources would markedly increase the purview of inferred cancer causation, revealing a full picture of the importance of diverse mechanisms behind the spectrum of genomic alterations fueling cancer evolution.

For an individual cancer patient, calculation of the relative cancer effect of diverse sources of mutation provides an estimate of how much each mutagenic process is responsible for an individual's cancer. From a public health perspective, these calculations constitute a bridge between molecular studies and long-standing epidemiological analyses that have associated behaviors (e.g., smoking) or professions (e.g., sun exposure) with cancer incidence. Public health intervention targeted at minimizing exposure to these actionable signatures would mitigate disease severity by preventing the accumulation of mutations that directly contribute to the cancer phenotype. Finally, our findings connect specific mutagenesis patterns and processes with

cancer, providing a "smoking gun" that can inform individuals as to why an instance of cancer happened—and have promise to play a significant role in demonstrating individual as well as group-level cause for legal recourse due to carcinogenic exposure (e.g. *54*).

The quantification of cancer effect attributable to specific sources of mutation has evident parallels to epidemiological results that assess the effect of risk factors on cancer causation (*55*). These epidemiological results often rely on correlation, and calculate an increase in the probability of cancer in relation to some behavior or exposure. Calculations of the relative cancer effect of diverse sources of mutation, in principle, directly relates the mutations driving tumorigenesis to mechanistic processes. However, multiple challenges impede their use at a population level in comparison to longstanding, well-crafted epidemiological studies: 1) conducting appropriate tumor sampling—most large tumor sequencing studies are sampled haphazardly, without reference to a distinct population, without stratification or even "random" sampling; 2) formulating an "apples to apples" quantitative mapping comparing proportions of effect to odds ratios; and 3) forming a discrete mapping of mutational signatures to mechanistic processes to epidemiological factors. These attributions of cause associated with COSMIC signatures are critical to our interpretations of these results, and range in surety from well-established (e.g., UV #7 and smoking #4), to presumptive (e.g. indirect damage from UV light #38).

Recent research has touched on a debate as to what extent "bad luck"—endogenous mutagenic processes that accumulate naturally with age—plays a role in the incidence of cancer arising in various tissues. Here, we addressed the question regarding the relative contributions of exogenous and endogenous sources of mutation to tumorigenesis by quantifying the extent that specific variants are driving tumorigenesis, and attributing the variants back to the mutational processes that originally fueled their creation. We found that signatures relating to aging processes (#1 and #5) were responsible for the majority of cancer effect in tumors of the brain (LGG, GBM) and tissues with large amounts of epithelial turnover (READ, COAD, STAD, UCEC). Other tumors whose cancer effects could largely be attributed to aging include PRAD (a tumor type strongly associated with age (*56*)), THCA (whose major single-nucleotide driver, BRAFV600E, is more likely to be caused by mutations associated with clock-like signature #5 than by mutations associated with other signatures), and PAAD. Several tumor types have large proportions of the cancer effect size directly attributable to mutational processes that are actionable, i.e., interventions could reduce the mutations in these tissues that are responsible for the cancer-causing variants. CESC, HNSC, and BLCA are largely driven by mutations attributed to virus-induced APOBEC activity, SKCM is largely driven by UV light exposure, and mutations responsible for increased proliferation and survival of cancerous cells within lung cancers trace back to smoking.

The importance of understanding the underlying sources of mutations that ultimately lead to cancer in each and every patient—whether they are endogenous or exogenous, and whether they come from sources that are actionable—is underscored by the remarkable successes of

anti-smoking interventions against carcinogenic exposures, which have saved many lives (*57*). In our study, some cancer types, such as OV, KIRC, ESCA, and LIHC, had a large proportion of cancer effect attributable to signatures with unknown etiology. As we gain greater insight into these mutational signatures and their diverse causative mechanisms, we may discover additional actionable mutational processes that can be mitigated by proactive public health interventions.

Acknowledgements

Members of the Townsend Lab provided stimulating discussions and helpful feedback on this research. Funding for this research was provided by NIH 1R01LM012487, NIH 1R01CA215900, NIH 5R01 CA231112, the Yale Cancer Biology Training program (NIH T32 CA193200/CA/NCI HHS/United States), and the Elihu Professorship endowed research funds.

Author Contributions

VLC and JPT designed the research. VLC executed all analyses. JDM provided computational tools and expertise. VLC and JPT wrote the manuscript. All authors reviewed the manuscript.

Declaration of Interests

JPT is currently engaged as a consultant for Black Diamond Therapeutics and Agios Pharmaceuticals. Other authors declare no potential conflicts of interest with the publication of this research.

References cited

1. P. C. Nowell, The clonal evolution of tumor cell populations. *Science*. **194**, 23–28 (1976).

2. J. A. Somarelli, H. Gardner, V. L. Cannataro, E. F. Gunady, A. M. Boddy, N. A. Johnson, J. N. Fisk, S. G. Gaffney, J. H. Chuang, S. Li, F. D. Ciccarelli, A. R. Panchenko, K. Megquier, S. Kumar, A. Dornburg, J. DeGregori, J. P. Townsend, Molecular Biology and Evolution of Cancer: From Discovery to Action. *Mol. Biol. Evol.* **37**, 320–326 (2020).

3. L. M. F. Merlo, J. W. Pepper, B. J. Reid, C. C. Maley, Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*. **6** (2006), pp. 924–935.

4. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7–30 (2020).

5. M. T. Smith, K. Z. Guyton, C. F. Gibbons, J. M. Fritz, C. J. Portier, I. Rusyn, D. M. DeMarini, J. C. Caldwell, R. J. Kavlock, P. F. Lambert, S. S. Hecht, J. R. Bucher, B. W. Stewart, R. A. Baan, V. J. Cogliano, K. Straif, Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ. Health Perspect.* **124**, 713–721 (2016).

6. J. L. Barnes, M. Zubair, K. John, M. C. Poirier, F. L. Martin, Carcinogens and DNA damage. *Biochem. Soc. Trans.* **46**, 1213–1224 (2018).

7. L. R. Yates, P. J. Campbell, Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).

8. M. Greaves, Evolutionary Determinants of Cancer. *Cancer Discovery*. **5** (2015), pp. 806–820.

9. E. A. Golemis, P. Scheet, T. N. Beck, E. M. Scolnick, D. J. Hunter, E. Hawk, N. Hopkins, Molecular mechanisms of the preventable causes of cancer in the United States. *Genes Dev.* **32**, 868–902 (2018).

10. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, PCAWG Mutational Signatures Working Group, G. Getz, S. G. Rozen, M. R. Stratton, PCAWG Consortium, The repertoire of mutational signatures in human cancer. *Nature*. **578**, 94–101 (2020).

11. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger,

D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature*. **500**, 415–421 (2013).

12. J. E. Kucab, X. Zou, S. Morganella, M. Joel, A. S. Nanda, E. Nagy, C. Gomez, A. Degasperi, R. Harris, S. P. Jackson, V. M. Arlt, D. H. Phillips, S. Nik-Zainal, A Compendium of Mutational Signatures of Environmental Agents. *Cell*. **177**, 821–836.e16 (2019).

13. S.-R. Hosseini, R. Diaz-Uriarte, F. Markowetz, N. Beerenwinkel, Estimating the predictability of cancer evolution. *Bioinformatics*. **35**, i389–i397 (2019).

14. S. Venkatesan, N. J. Birkbak, C. Swanton, Constraints in cancer evolution. *Biochemical Society Transactions*. **45** (2017), pp. 1–13.

15. D. E. Brash, J. A. Rudolph, J. A. Simon, A. Lin, G. J. McKenna, H. P. Baden, A. J. Halperin, J. Pontén, A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 10124–10128 (1991).

16. G. P. Pfeifer, How the environment shapes cancer genomes. *Curr. Opin. Oncol.* **27**, 71–77 (2015).

17. G. P. Pfeifer, M. F. Denissenko, M. Olivier, N. Tretyakova, S. S. Hecht, P. Hainaut, Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*. **21**, 7435–7451 (2002).

18. S. L. Poon, J. R. McPherson, P. Tan, B. T. Teh, S. G. Rozen, Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Med.* **6**, 24 (2014).

19. R. Segovia, A. S. Tam, P. C. Stirling, Dissecting genetic and environmental mutation signatures with model organisms. *Trends Genet.* **31**, 465–474 (2015).

20. J. E. Grolleman, M. Díaz-Gay, S. Franch-Expósito, S. Castellví-Bel, R. M. de Voer, Somatic mutational signatures in polyposis and colorectal cancer. *Mol. Aspects Med.* **69**, 62–72 (2019).

21. C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes,

K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, M. R. Stratton, Patterns of somatic mutation in human cancer genomes. *Nature*. **446**, 153–158 (2007).

22. V. L. Cannataro, J. P. Townsend, Neutral theory and the somatic evolution of cancer. *Mol. Biol. Evol.* (2018), doi:10.1093/molbev/msy079.

23. R. C. Poulos, Y. T. Wong, R. Ryan, H. Pang, J. W. H. Wong, Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. *PLoS Genet.* **14**, e1007779 (2018).

24. S. L. Ostrow, R. Barshir, J. DeGregori, E. Yeger-Lotem, R. Hershberg, Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS Genet.* **10**, e1004239 (2014).

25. M. Shpak, J. Lu, An Evolutionary Genetic Perspective on Cancer Biology. *Annu. Rev. Ecol. Evol. Syst.* **47**, 25–49 (2016).

26. Z.-M. Zhao, B. Zhao, Y. Bai, A. Iamarino, S. G. Gaffney, J. Schlessinger, R. P. Lifton, D. L. Rimm, J. P. Townsend, Early and multiple origins of metastatic lineages within primary tumors. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2140–2145 (2016).

27. J. Van den Eynden, E. Larsson, Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric. *Front. Genet.* **8** (2017), doi:10.3389/fgene.2017.00074.

28. I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, P. J. Campbell, Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. **171**, 1029–1041 (2017).

29. D. Temko, I. P. M. Tomlinson, S. Severini, B. Schuster-Böckler, T. A. Graham, The effects of mutational processes and selection on driver mutations across cancer types. *Nat. Commun.* **9**, 1857 (2018).

30. V. L. Cannataro, S. G. Gaffney, J. P. Townsend, Effect Sizes of Somatic Mutations in Cancer. *J. Natl. Cancer Inst.* **110**, 1171–1177 (2018).

31. R. Doll, A. B. Hill, Smoking and carcinoma of the lung; preliminary report. *Br. Med. J.* **2**, 739–748 (1950).

32. L. A. Loeb, C. C. Harris, Advances in chemical carcinogenesis: a historical review and

prospective. *Cancer Res.* **68**, 6863–6872 (2008).

33. C. Tomasetti, B. Vogelstein, Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*. **347**, 78–81 (2015).

34. C. Tomasetti, L. Li, B. Vogelstein, Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*. **355**, 1330–1334 (2017).

35. S. Wu, S. Powers, W. Zhu, Y. A. Hannun, Substantial contribution of extrinsic risk factors to cancer development. *Nature*. **529**, 43–47 (2016).

36. N. A. Ashford, P. Bauman, H. S. Brown, R. W. Clapp, A. M. Finkel, D. Gee, D. B. Hattis, M. Martuzzi, A. J. Sasco, J. B. Sass, Cancer risk: Role of environment. *Science*. **347**, 727–727 (2015).

37. C. Wild, P. Brennan, M. Plummer, F. Bray, K. Straif, J. Zavadil, Cancer risk: role of chance overstated. *Science*. **347** (2015), p. 728.

38. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, C. Swanton, DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).

39. M. Petljak, L. B. Alexandrov, Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*. **37** (2016), pp. 531–540.

40. L. B. Alexandrov, M. Zhivagui, Mutational Signatures and the Etiology of Human Cancers. *Reference Module in Biomedical Sciences* (2018), , doi:10.1016/b978-0-12-801238-3.65046-8.

41. L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, M. R. Stratton, Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

42. R. Dubrow, A. S. Darefsky, Demographic variation in incidence of adult glioma by subtype, United States, 1992-2007. *BMC Cancer*. **11** (2011), doi:10.1186/1471-2407-11-325.

43. P. Rawla, Epidemiology of Prostate Cancer. *World J. Oncol.* **10**, 63–89 (2019).

44. M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A.

Gordenin, S. Sunyaev, E. S. Lander, G. Getz, Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. **499**, 214–218 (2013).

45. M. Fousteri, L. H. F. Mullenders, Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res.* **18**, 73–84 (2008).

46. J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, S. R. Sunyaev, Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).

47. M. P. Ramakodi, R. J. Kulathinal, Y. Chung, I. Serebriiskii, J. C. Liu, C. C. Ragin, Ancestral-derived effects on the mutational landscape of laryngeal cancer. *Genomics*. **107**, 76–82 (2016).

48. N. D. Montgomery, S. R. Selitsky, N. M. Patel, D. Neil Hayes, J. S. Parker, K. E. Weck, Identification of Germline Variants in Tumor Genomic Sequencing Analysis. *The Journal of Molecular Diagnostics*. **20** (2018), pp. 123–125.

49. W. Sun, P. Bunn, C. Jin, P. Little, V. Zhabotynsky, C. M. Perou, D. N. Hayes, M. Chen, D.-Y. Lin, The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Res.* **46**, 3009–3018 (2018).

50. E. A. Mroz, A. D. Tward, R. J. Hammon, Y. Ren, J. W. Rocco, Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the Cancer Genome Atlas. *PLoS Med.* **12**, e1001786 (2015).

51. E. Laconi, F. Marongiu, J. DeGregori, Cancer as a disease of old age: changing mutational and microenvironmental landscapes. *Br. J. Cancer*. **122**, 943–952 (2020).

52. L. A. Liggett, J. DeGregori, Changing mutational and adaptive landscapes and the genesis of cancer. *Biochim. Biophys. Acta Rev. Cancer*. **1867**, 84–94 (2017).

53. G. Macintyre, T. E. Goranova, D. De Silva, D. Ennis, A. M. Piskorz, M. Eldridge, D. Sie, L.-A. Lewsley, A. Hanif, C. Wilson, S. Dowson, R. M. Glasspool, M. Lockley, E. Brockbank, A. Montes, A. Walther, S. Sundar, R. Edmondson, G. D. Hall, A. Clamp, C. Gourley, M. Hall, C. Fotopoulou, H. Gabra, J. Paul, A. Supernat, D. Millan, A. Hoyle, G. Bryson, C. Nourse, L. Mincarelli, L. N. Sanchez, B. Ylstra, M. Jimenez-Linan, L. Moore, O. Hofmann, F. Markowetz, I. A. McNeish, J. D. Brenton, Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).

54. S. G. Lee, Proving Causation With Epidemiological Evidence in Tobacco Lawsuits. *J. Prev. Med. Public Health*. **49**, 80–96 (2016).

55. K. D. Shield, D. Maxwell Parkin, D. C. Whiteman, J. Rehm, V. Viallon, C. M. Micallef, P. Vineis, L. Rushton, F. Bray, I. Soerjomataram, Population Attributable and Preventable Fractions: Cancer Risk Factor Surveillance, and Cancer Policy Projection. *Current*

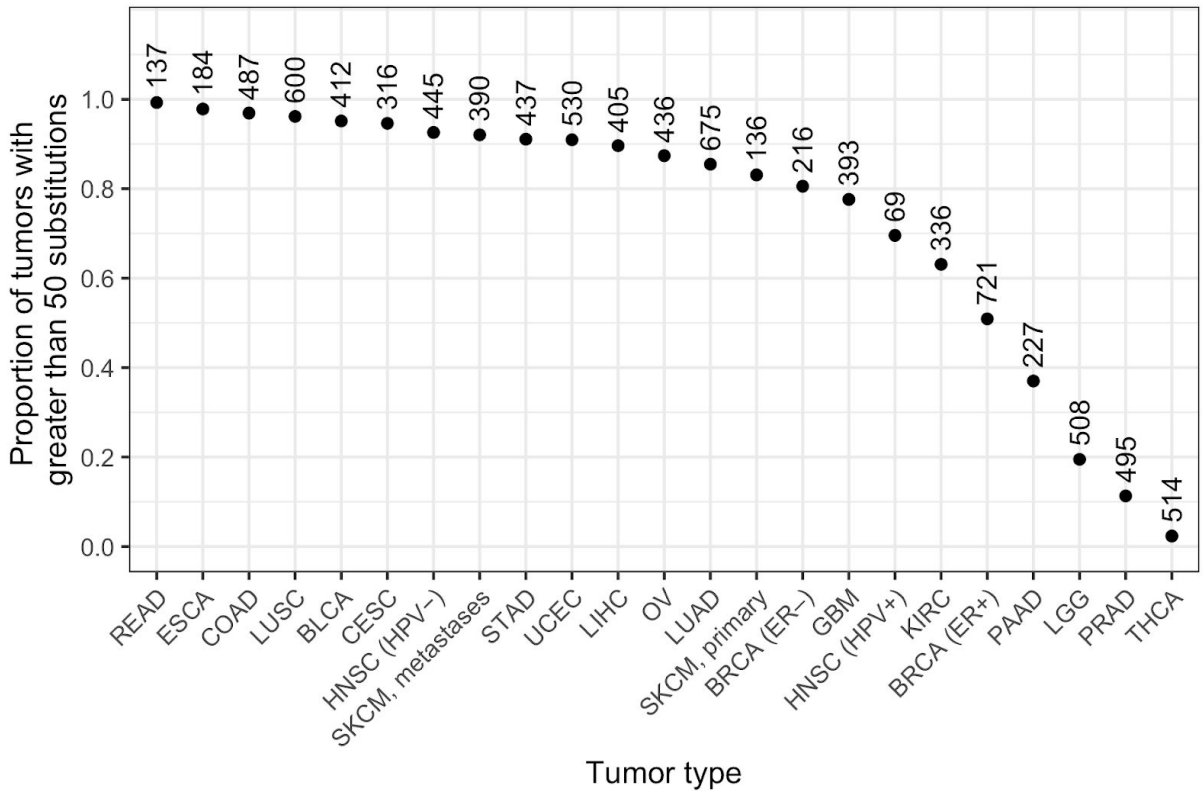
Epidemiology Reports. **3**, 201–211 (2016).

56. D. G. Bostwick, H. B. Burke, D. Djakiew, S. Euling, S.-M. Ho, J. Landolph, H. Morrison, B. Sonawane, T. Shifflett, D. J. Waters, B. Timms, Human prostate cancer risk factors. *Cancer*. **101**, 2371–2490 (2004).

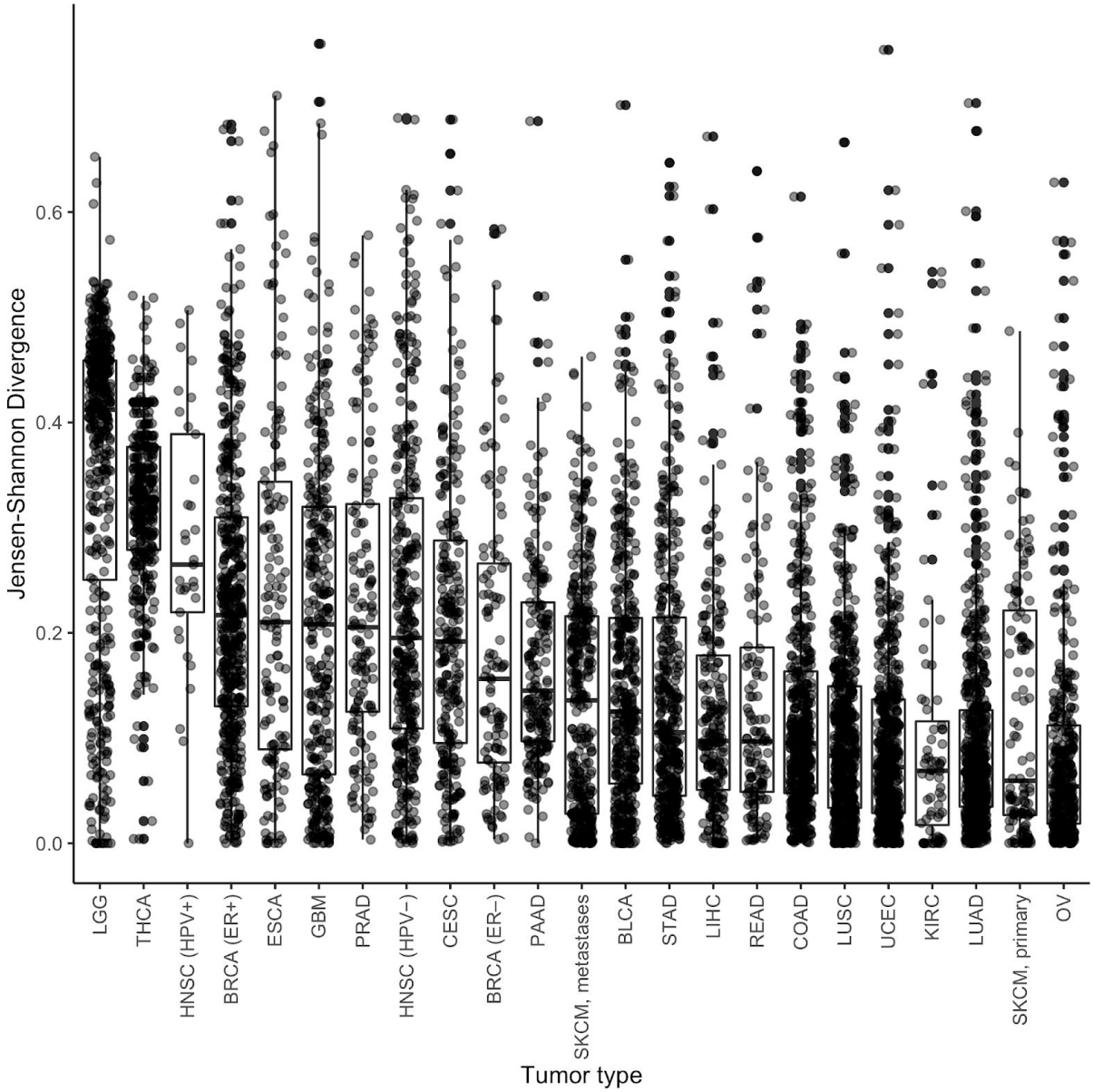
57. T. R. Holford, R. Meza, K. E. Warner, C. Meernik, J. Jeon, S. H. Moolgavkar, D. T. Levy, Tobacco control and the reduction in smoking-related premature deaths in the United States, 1964-2012. *JAMA*. **311**, 164–171 (2014).

Supplemental Table S1: Yale-Gilead sequenced tumors that received chemotherapies and were removed from this analysis

PY-11T	yukatTM	yucarTM	yurocTM
PY-3T	yuklabTM	yucasTM	yurogTM
PY-4T	yuksiTM	yucoveTM	yurolTM
PY-6T	yulanTM	yudabTM	yurosTM
PY-8T	yulyoTM	yudateTM	yusariTM
yucavT	yumarTM	yudedeTM	yuscaTM
yutucoTM	yumerTM	yufitTM	yusimTM
yuvonTM	yumezTM	yuflaTM	yusivTM
yuzestTM4	yumutTM	yugaspTM	yusmiTM
yualeTM	yunuffTM	yugemaTM	yusopTM
yuaveyTM	yuomegaTM	yugismoTM	yuspoTM
yubanTM	yupaerTM	yugoeTM	yuswiTM
yubigTM	yuplaTM	yugurtTM	yutreTM
yublastTM	yurayTM	yuhuyTM3	yuvemeTM
yubowTM3	yuredTM	yujubeTM2	yuzestTM3
yubuneTM	yuridaTM2	yurifTM	



Supplementary Figure S1: The proportions of tumors with greater than 50 substitutions in 23 tumor types. Numbers above the points indicate the total number of tumors in the data subset.



Supplemental Figure S2. Box plots by tumor type of the Jensen-Shannon Divergence between the proportional input of each mutagenic source and the proportion that each signature contributed to the total cancer effect in each tumor, for 23 cancer types: brain lower-grade glioma, thyroid carcinoma, HPV^+ head and neck squamous-cell carcinoma, ER^+ breast invasive carcinoma, esophageal carcinoma, glioblastoma multiforme, prostate adenocarcinoma, HPV^- head and neck squamous-cell carcinoma, cervical squamous cell carcinoma and endocervical adenocarcinoma, ER^- breast invasive carcinoma, pancreatic adenocarcinoma, skin cutaneous melanoma (metastatic), bladder urothelial carcinoma, stomach adenocarcinoma, liver hepatocellular carcinoma, rectum adenocarcinoma, colon adenocarcinoma, lung squamous-cell carcinoma,uterine corpus endometrial carcinoma,kidney renal clear cell carcinoma, lung adenocarcinoma, skin cutaneous melanoma (primary), ovarian serous cystadenocarcinoma.