

COMMENTARY

Effect Sizes of Somatic Mutations in Cancer

Vincent L. Cannataro, Stephen G. Gaffney, Jeffrey P. Townsend

See the Notes section for the full list of authors' affiliations.

Correspondence to: Jeffrey P. Townsend, PhD, Department of Biostatistics, Yale School of Public Health, 135 College Street, New Haven, CT 06510 (e-mail: Jeffrey.Townsend@Yale.edu).

Abstract

A major goal of cancer biology is determination of the relative importance of the genetic alterations that confer selective advantage to cancer cells. Tumor sequence surveys have frequently ranked the importance of substitutions to cancer growth by *P* value or a false-discovery conversion thereof. However, *P* values are thresholds for belief, not metrics of effect. Their frequent misuse as metrics of effect has often been vociferously decried, even in cases when the only attributable mistake was omission of effect sizes. Here, we propose an appropriate ranking—the cancer effect size, which is the selection intensity for somatic variants in cancer cell lineages. The selection intensity is a metric of the survival and reproductive advantage conferred by mutations in somatic tissue. Thus, they are of fundamental importance to oncology, and have immediate relevance to ongoing decision making in precision medicine tumor boards, to the selection and design of clinical trials, to the targeted development of pharmaceuticals, and to basic research prioritization. Within this commentary, we first discuss the scope of current methods that rank confidence in the overrepresentation of specific mutated genes in cancer genomes. Then we bring to bear recent advances that draw upon an understanding of the development of cancer as an evolutionary process to estimate the effect size of somatic variants leading to cancer. We demonstrate the estimation of the effect sizes of all recurrent single nucleotide variants in 22 cancer types, quantifying relative importance within and between driver genes.

Since the advent of whole-exome and whole-genome sequencing of tumor tissues, studies of somatic mutations have revealed the underlying genetic architecture of cancer (1), producing lists of statistically significantly mutated genes whose ordering implies their relative importance to tumorigenesis and cancer development. Typically, differentiation of selected mutations from neutral mutations is performed by quantifying the overrepresentation of mutations within specific genes in tumor tissue relative to normal tissue, and disproportionate prevalence of somatic mutations in a gene has been taken as *prima facie* evidence of a causative role for that gene. Two quantifications have implicitly ordered the importance of discovered cancer “driver” genes: the prevalence of the mutation among tumor tissues sequenced from that tumor type (2,3), the statistical significance (*P* value) of the disproportionality of mutation frequency (4), or both (1). Versions of these metrics have shifted from simple ranks by mutation prevalence in a tumor population (5–7) to calculation of statistical significance of mutation prevalence over genome-wide context-specific background

mutation rates (8–10), to ratios of the prevalence of nonsynonymous and synonymous mutations (11), to *P* values based on a gene-specific mutation rate and a diversity of genomic data (12). Although the approaches used to calculate *P* values have become more sophisticated, *P* values are not an appropriate metric for quantifying the vital role of genes or their mutations to tumorigenesis and cancer development, as *P* values are thresholds for belief (13) and not metrics of effect (14,15). Failure to report effect sizes is a persistent issue in the scientific and biomedical literature that can massively misdirect research and health-care decision making (16–21).

Although prevalence of a somatic mutation in a cancer type has important consequences for biomarker studies (22) and identification of the therapeutic population for a targeted therapy (23), there is only a correlative—rather than causal—link between prevalence of mutation and its contribution to tumorigenesis and cancer development. The lack of causal linkage is easily seen by considering the mutated genes that, despite their high prevalence in tumor populations, are

Received: March 27, 2018; Revised: July 20, 2018; Accepted: August 24, 2018

© The Author(s) 2018. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

universally regarded as false positives. For example, the gene *TTN* encodes a structural protein of striated muscle. Because it is long, because it replicates relatively late in the DNA synthesis phase of the cell cycle, and because it is inaccessible to transcription-mediated repair in nonmuscle tissues, it has a high mutation rate and is frequently mutated in non-sarcoma cancer tissues. Granted, *TTN* is an extreme example—sometimes showing up at the top of lists ordered by prevalence of genic mutation (11,12)—but it exemplifies the problem with using substitution prevalence as a proxy for importance. Any consideration of whether mutated genes are contributing to tumorigenesis and cancer development—or of the degree to which they are contributing—must address the issue of their underlying mutation rate.

The appearance of such biologically implausible genes in ranked mutation lists prompted the development of increasingly sophisticated statistical approaches designed to “weed out” false positives via calculation of a *P* value that accounted for gene length and background mutation rate. The classical evolutionary biology approach is to use the frequency of synonymous site mutations in each gene as a proxy for mutation rate. As in the divergence of species, synonymous site mutations are presumably neutral (or nearly so) to the success of cancer lineages during the divergence from normal to resectable tumor. As the number of mutations observed in a given gene is typically much smaller in the somatic evolution of cancer than is observed in the divergence between most species, use of synonymous sites within a single gene leads to many genes with zero synonymous mutations in most cancers, and an ineffective calculation of *P* value. Alternative approaches currently in use obtain a reasonably robust estimate of genic mutation rate by incorporating correlates of mutation rate among genes such as gene expression levels, chromatin states, and replication timing. These approaches are largely successful at excluding known false positives (2,12).

The sample size of tumors varies among genomic tumor surveys, posing a problem for comparison of *P* values within or between cancer types (14,15). An even more serious issue with using *P* values for ranking genes or mutations arises from the same source that obviates use of genic mutation prevalence: the confounding effect of mutation rate. Because the “sample size” of overall mutations in a gene is dictated by the genic mutation rate, it is much easier for genes with high mutation rates not only to reach high genic prevalence, but also to reach statistical significance despite small effect sizes. Although approaches accounting for genic mutation rates will eliminate false positives (2,12), and the *P* value will serve to exclude genes like *TTN* that have no role in tumorigenesis and cancer development, the rank order by *P* value of genes that do have a role in tumorigenesis and cancer development will remain highly affected by mutation rate. Genes with higher mutation rate will (correctly) be more likely to achieve statistical significance, and thus will appear deceptively high on a ranked list whose ordering suggests importance in tumorigenesis and cancer development.

Because genic mutation prevalence and *P* value inadequately capture importance to tumorigenesis and cancer development, we aimed to derive a more appropriate method. To evaluate the relative importance of mutations in diverse cancer types to tumorigenesis and cancer development, we adapted insights from classical evolutionary theory, informed by an understanding of the development of cancer as an evolutionary process (24,25). By applying an evolutionary concept, the importance of a mutation to tumorigenesis and cancer development

can be quantified by its “selective effect” on the cancer lineage. The appropriateness of this metric is fairly easy to recognize. Mutations are a major source of variation contributing to tumorigenesis, yet we do not conduct genomic tumor sequence surveys to discover neutral mutation rates: we conduct them to determine which mutations spread within cancer tissues because of the effects of mutations on proliferation and survival. Mutation rate is a confounding phenomenon—when it is high, it also increases allelic fraction and the probability of substitution.

Because silent site substitutions and other correlates of baseline mutation rate provide a means to estimate silent mutation rate independently from the impact of natural selection within the tumor, selection intensities can also be estimated, providing the effect sizes of each mutation. In accordance with population genetic theory, the rate at which neutral mutations arise and the rate at which they fix as substitutions are equivalent, and non-neutral mutations arise at a consistent rate. Therefore, the rate that we expect to see specific nucleotide variants within tumors in the absence of selection can be calculated by quantifying the rate of silent mutation at the resolution of genes (12), and adjusting this rate at the nucleotide level so as to be consistent with the average trinucleotide mutational signatures observed in each tumor type (26,27). The flux of substitutions among tumors at a nucleotide site that occurs above the baseline silent rate represents the intensity of selection on that mutation within the tumor population. These selection intensities quantify the survival and proliferative advantage conferred by variants, facilitating comparisons of the relative importance of drivers among and within tumor types.

Methods

Data Acquisition and Preprocessing

Data were obtained from the National Cancer Institute’s The Cancer Genome Atlas (NCI TCGA) database (28) and from the Yale-Gilead collaboration (Supplementary Table 1, available online). All TCGA data were first converted to hg19 coordinates using the liftOver function of the R package rtracklayer (29) to ensure compatibility with all software used. Nucleotide variants that were only one or two positions apart in the same tumor sample were removed from the analysis to ensure that we analyzed only single-nucleotide variants. We used tissue-specific mutational covariates to inform mutation rate calculations and thus used all available data on gene expression levels from The Genotype-Tissue Expression (GTEx) portal (30) and the Cancer Cell Line Encyclopedia (CCLE) database (31), replication timing from ReplicationDomain (32), and chromatin state from the RoadMap Epigenomics Project (33). The specific files used for each cancer type are listed in Supplementary Table 2 (available online). Head and neck squamous cell carcinoma (HNSC) tumors with variant data obtained from the NCI TCGA database were designated as positive for human papillomavirus (HPV) if they contained greater than 100 HPV RNA viral transcript reads per hundred million (34), and one of the 17 tumors obtained from Hedberg et al. (35) was determined to be HPV-associated by HPV in situ hybridization and p16 immunohistochemistry (36). Breast invasive carcinoma (BRCA) tumors were designated as estrogen receptor-positive and -negative as described in the TCGA GDAC Firehose (37).

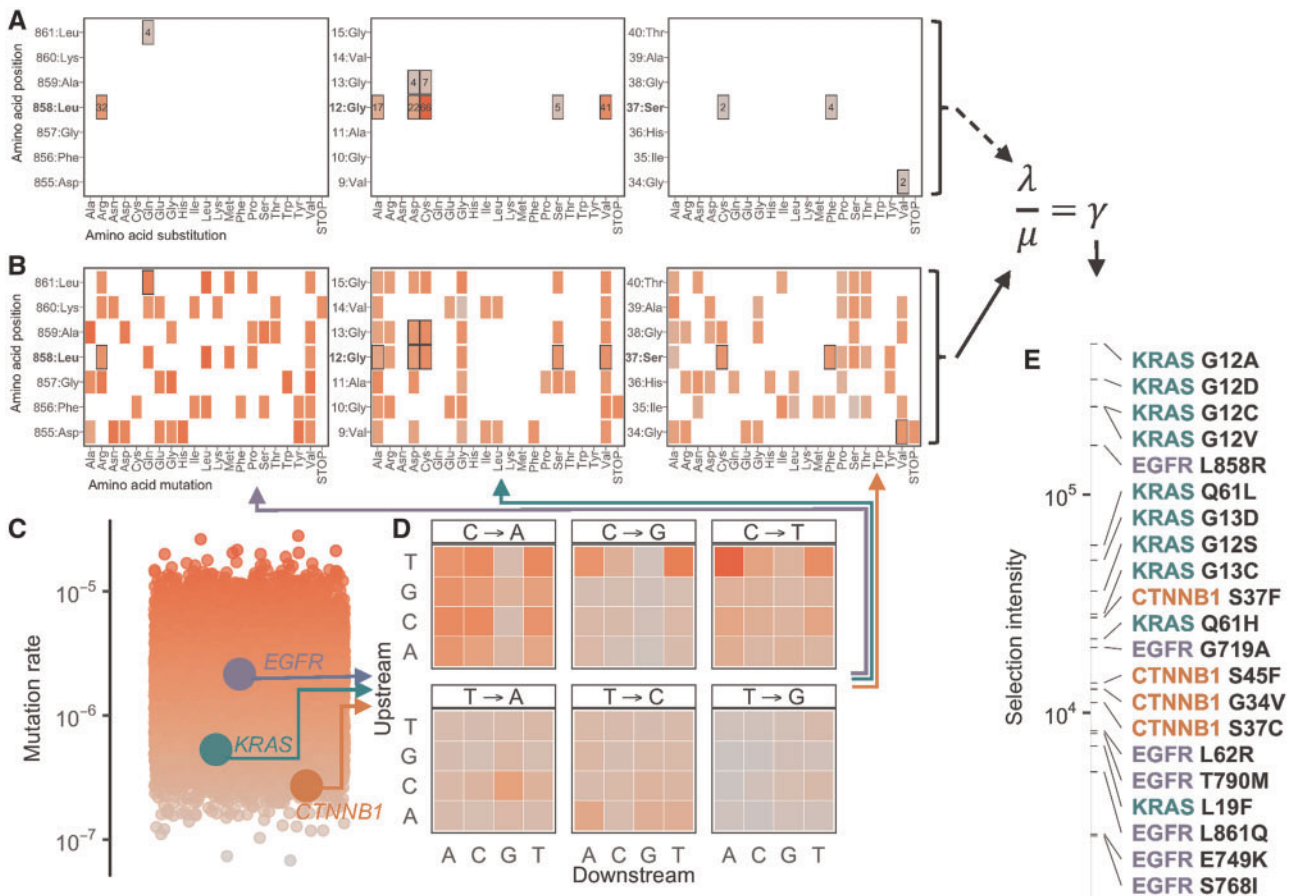


Figure 1. Deconvolution of prevalence by mutation rates to yield selection intensity for recurrent amino acid mutations within three oncoproteins caused by single-nucleotide changes in lung adenocarcinoma. **A)** Observed substitution rates—after correction under an assumption of complete intragenic epistasis—are divided by **(B)** the expected substitution rates in the absence of selection. Expected substitution rates in the absence of selection are calculated as **(C)** the average per-site synonymous mutation rate of the gene, normalized for **(D)** the average weight of trinucleotide mutational signature burden for that tissue. The quotient of observed to expected numbers of substitutions is **(E)** the selection intensity, or cancer effect size, of variants.

Effect Size Calculation

Gene-level mutation rates were calculated using the maximum likelihood estimate for the expected number of synonymous mutations in a gene with the *dndscv* package and tissue-specific covariate files. These rates were converted to the expected average mutation rate per silent site by dividing each rate by the number of potential silent mutations in each gene. The mutation rate for sites across each gene was then calculated as the mutation rate of each trinucleotide context, weighted such that the average mutation rate among all sites equaled the average site-specific mutation rate for that gene. Trinucleotide-specific mutation rates were calculated using *deconstructSigs*, and were derived from all signatures present in each tumor sample. This mutation rate is presumably equivalent to the rate at which we would expect to observe mutations in tumor sequence if there were no selection. It can then be compared to the rate of which mutations are actually observed, ie, the substitution rate, after correction of the flux of substitutions given that we can only observe one substitution per site per normal-tissue to tumor-tissue comparison, and enforcement of a specification that only one substitution per gene per normal-tissue to tumor-tissue comparison is under selection (Figure 1, Supplementary Methods, available online). The intensity of selection, or cancer effect size, per variant is then equivalent to the substitution rate divided by the mutation rate. A

software package containing the tools to calculate cancer effect size, *cancereffectsizeR*, is available for free download and usage at <https://github.com/Townsend-Lab-Yale/cancereffectsizeR>.

Statistical Analysis

The Q values reported in figures are calculated by *dndscv* as described by Martincorena et al. (2), and use multiple testing adjustment via the Benjamini-Hochberg false discovery rate. Confidence intervals of all effect sizes were calculated by bootstrapping the tumor samples 1000 times. Confidence intervals are reported in Supplementary Table 3 (available online).

Results and Discussion

To calculate the cancer effect sizes of all recurrent single-nucleotide variants in 22 cancer types, we compared the rate of observed substitutions to the rate that substitutions would be expected to arise in the absence of selection (38) in data that either were obtained from TCGA projects downloaded from the National Cancer Institute's Genomic Data Commons (28) (Supplementary Table 2, available online) or were derived from whole exome sequencing of tumors as part of a collaboration with Gilead Sciences (Supplementary Table 1, Methods,

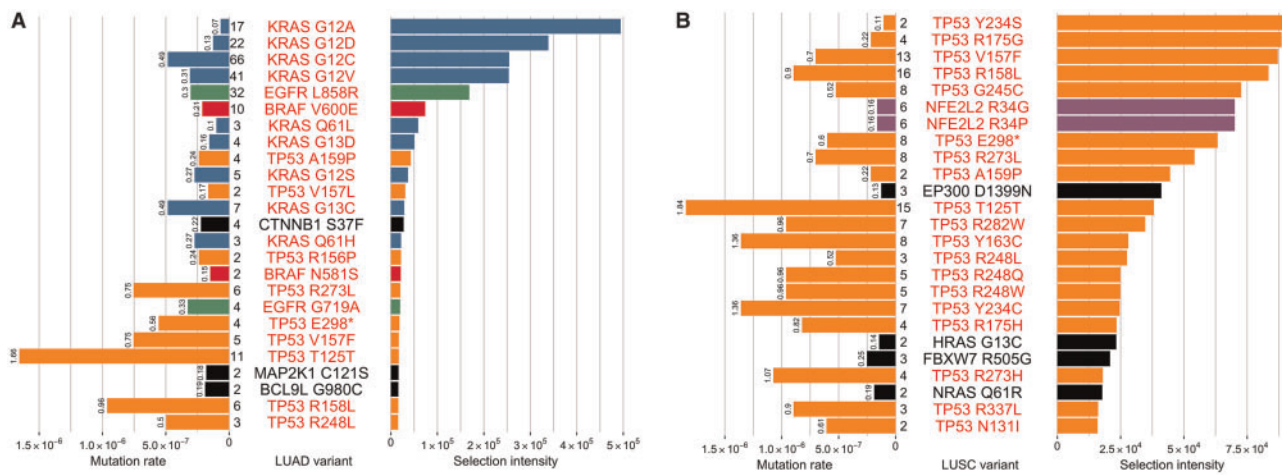


Figure 2. The 25 single-nucleotide variants (red = $dndscv\ Q < 0.05$; black = $dndscv\ Q \geq 0.05$) with the highest selection intensity from (A) 675 LUAD tumors and (B) 600 LUSC tumors, within the expertly curated COSMIC list of drivers (39), ranked by selection intensity (bar, uniquely colored by gene identity, to the right of variant name), determined by deconvolving prevalence (number to the left of the variant name), by mutation rate (bar, uniquely colored by gene identity, to the left of prevalence, and rate $\times 10^{-6}$, to the left of the bar). The asterisk corresponds to a nonsense mutation. LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma.

available online). This estimation of tissue-specific site-specific mutation rates accounts for mutagenic exposure at the gene level and mutational patterns at the trinucleotide level; the estimation is important because these rates can vary substantially between tissue types for the same mutation. For instance, the rate we would expect a mutation creating a BRAF V600E variant in a cancer cell is an order of magnitude different between lung adenocarcinoma and skin cutaneous melanoma (Supplementary Table 3, available online). Deconvolution by this mutation rate reveals a wide distribution of selection intensities in somatic tumors—selection intensities that vary considerably from tumor type to tumor type, even among tumors that arise from tissues exposed to similar mutational pressures. For instance, even though lung tissues encounter similar mutagens and hence mutations arise through similar mutational signatures, lung adenocarcinoma and lung squamous cell carcinoma variants conferring the largest effect sizes are markedly different among the two cancers (Figure 2). Moreover, within a tumor type, the selection intensity decouples the effects of mutation rate on frequency: Even though EGFR L858R variants are approximately twice as prevalent as KRAS G12A variants in lung adenocarcinoma tumors, KRAS G12A is estimated to have a higher effect size, because of its much lower baseline mutation rate. Furthermore, the relative selection intensity among variants within a single gene reveals whether “hotspots” of somatic variation are purely driven by mutational processes or, in contrast, are under differential selective pressure. For instance, the high prevalence of KRAS G12C relative to other KRAS position 12 substitutions within lung adenocarcinoma is often attributed to the high G \rightarrow T nucleotide mutation rate found in lung cancer tissue caused by tobacco smoke (38,40), and is not typically attributed to an increased proliferative advantage conferred by this particular mutation relative to other KRAS position 12 mutations. Indeed, we calculate remarkably similar estimates of the effect sizes of KRAS G12 variants—despite large differences in substitution rates (mutation prevalence).

This calculation yielded cancer effect sizes for all fixed substitutions (Supplementary Figure 1, Supplementary Table 3, available online) that quantify contribution to the cancer phenotype within 22 tumor types. As should be expected,

substitutions in previously described “hotspots” (41,42) had statistically significantly higher selection intensities than substitutions outside of previously described hotspots (one-sided Wilcoxon rank sum test, $P < 2.2 \times 10^{-16}$). Several common known oncogenic substitutions, such as BRAF V600E in primary skin cutaneous melanoma and EGFR L858R in lung adenocarcinoma (43,44), and substitutions in known tumor suppressor genes, such as APC in rectum adenocarcinoma and TP53 in HPV⁻ HNSC, are highly selected. These genes are also typically determined as statistically significantly mutated [eg, by $dndscv$ (2)]. However, genes that have been determined to be statistically significantly altered in cancer are well dispersed across a large range of site-specific cancer effect sizes (Figure 3), illustrating how discrepant gene-specific Q values are with site-specific cancer effect sizes. Several substitutions within genes that are not statistically significantly overmutated are interspersed among more prevalent substitutions within genes that are estimated to be statistically significantly mutated, for instance, Mastermind-like3 (MAML3) G1073A in rectum adenocarcinoma, a protein that is known to bind to and stabilize the DNA-binding complex of the Notch intracellular domain (45) and ubiquitin ligase Cullin-3 (CUL3) M299R in prostate adenocarcinoma, which was only recently implicated as a driver of prostate cancer by an even higher sample-size tumor sequence analysis (3). The cancer effect size not only captures the distinct selective pressures underlying the development of different cancer types, it also captures the distinct selective pressures among tumors of the same type but different ontogeny. For instance, HPV⁺ and HPV⁻ HNSC and estrogen receptor-positive and estrogen receptor-negative BRCA each have distinct variants of high cancer effect size (notably, variants within TP53 rank among the highest effect sizes within HNSC and BRCA tumors negative for HPV and negative for ER, respectively; Supplementary Figure 1, available online). These distinct selective effects of somatic mutations likely reflect divergent evolutionary landscapes driven by epistatic interactions within each subtype (46,47). In contrast, uterine corpus endometrial carcinoma and colon adenocarcinoma tumor types analyzed within subsets classified by microsatellite instability—where the mutation rate is elevated but the underlying selective pressures

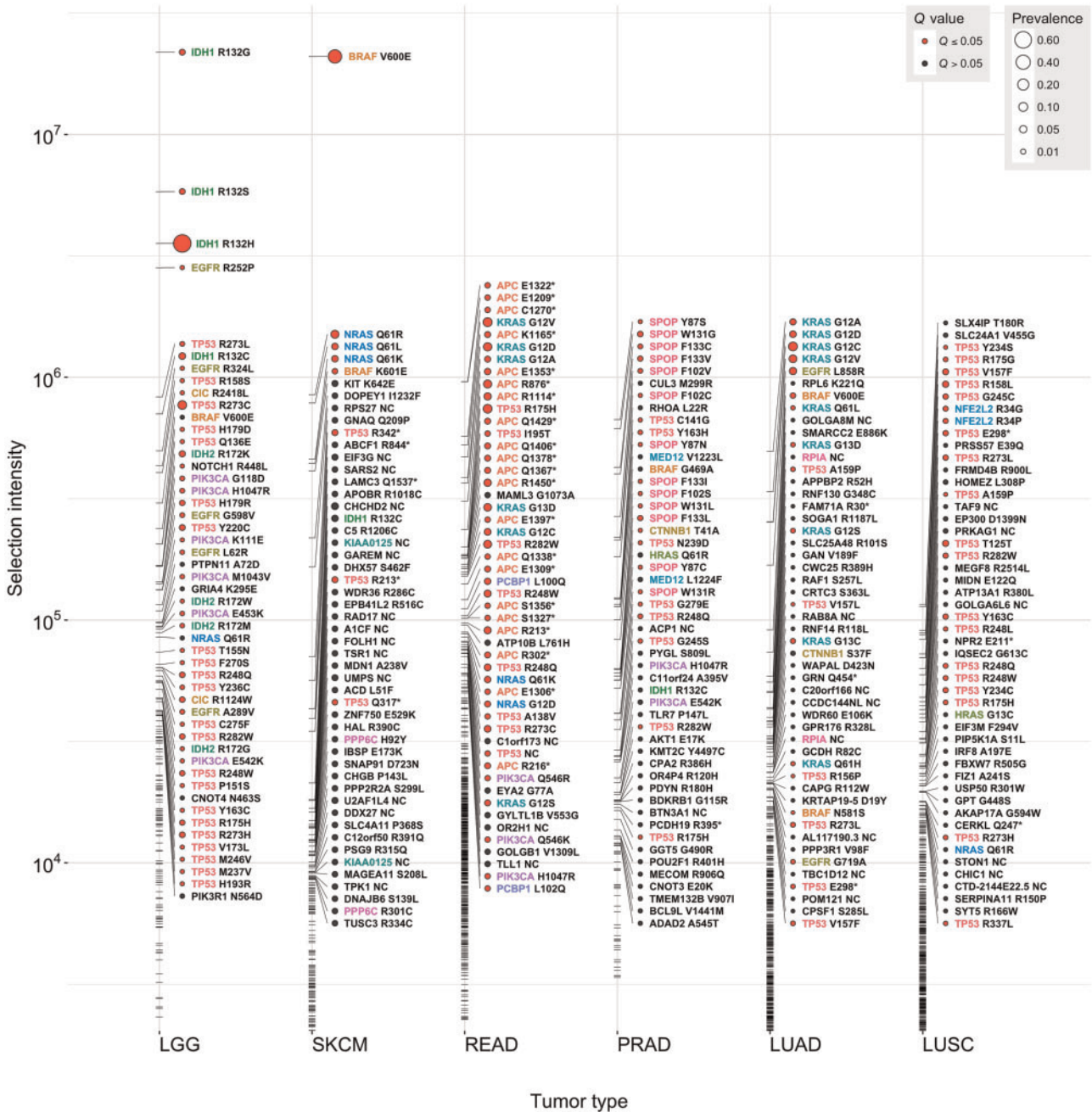


Figure 3. Cancer effect sizes of recurrent somatic substitutions in six of the 22 cancer types analyzed. Effect sizes greater than $1 \times 10^{3.5}$ are indicated by ticks along the tumor-type axes. The highest 50 effect sizes are labeled within each tumor. Names of genes that have more than one mutation listed within or between tumors are uniquely colored. Genes deemed statistically significantly overburdened with nonsynonymous mutation (2) are depicted by a red circle next to variant names, and the prevalence of each substitution is represented by the size of this circle. NC refers to a non-coding single-nucleotide variant outside an exon (eg, 5' or 3' UTRs). 108 LUAD, 108 LUSC, and 28 SKCM tumors from the Yale-Gilead collaboration are included in the plot. LGG = brain lower grade glioma; SKCM = skin cutaneous melanoma (primary); READ = rectum adenocarcinoma; PRAD = prostate adenocarcinoma; LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma.

driving tumorigenesis are similar—exhibit little difference in the relative selection intensity of mutations.

Here we have argued that frequencies of mutation and *P* values are not an appropriate metric for the importance of somatic variants in tumor growth. We derive the appropriate metric: the cancer effect size of mutations. This effect size—quantifying the intensity of selection on mutations in cancer cells in patients—conveys the replicative and survival benefit conferred by genetic variants, and therefore is a direct metric of the contribution of a

variant to the cancer phenotype. Using our approach, we estimated the effect size of all recurrent single-nucleotide variants in 22 cancer types, reevaluating their importance across single-nucleotide variant effect size to cancer in disparate tumor types projected to account for approximately 82% of all new cancer cases within the United States in 2017 (48).

Current approaches using conservative *P* values are particularly underpowered to detect genes that are of high importance to tumorigenesis and cancer development in some cancer cases

because the site or sites conveying the relevant phenotype are mutated at low rates. For instance, FBXW7 R505G was estimated to have the highest selection intensity in HPV⁺ HNSC, BRAF V600E was estimated to have the seventh highest selection intensity in low-grade glioma, and BRAF G469A was estimated to have the 12th highest selection intensity in prostate adenocarcinoma, yet these two genes were not classified as statistically significantly mutated at the gene level within these three cancer types. Mutations within these two well-known oncogenes were estimated to confer large effect sizes, and these genes were determined to be statistically significantly mutated in other cancer types, yet their importance in patients who have these rarer somatic variants within HPV⁺ HNSC, low-grade glioma, and prostate adenocarcinoma has been poorly illuminated by gene-wide analyses of statistically significant mutation burden across patients. Thus, calculating the cancer effect size at the level of single nucleotide variants identifies drivers with an underwhelming, low prevalence that potentially exert a very large effect in different tumor types.

Identification of these low-prevalence, high-effect drivers is increasingly important as precision targeting of therapeutics becomes increasingly integrated into clinical trial design. Targeted therapies are often developed for, and necessarily tested in, a single tumor type with the targeted variant at high frequency. However, the same variant often exists at lower prevalence in other tumor types. Quantifications of the cancer effect size can guide the selection and design of clinical trials to target small populations that can benefit from targeted therapeutics developed for other cancer types. Targets for therapies that were originally developed for variants at high prevalence in one tumor type are expected to be effective when treating the same variant in a secondary tumor type if the variant has similarly high selection intensity in the second tumor type, even if it is at low prevalence in the newly considered tumor type (49).

The effect sizes of somatic mutations in cancer should play key roles in clinical decision making, providing crucial insight into the relative upper limits of the efficacy of precision-targeted treatments. For instance, the relative ranking of the effect sizes of the somatic variants within a tumor indicates the variants that, when targeted, would have the largest predicted effect on tumor progression. These predictions of evolutionary effect given targeted therapy complement preclinical research on drug efficacy, toxicity, and epistatic interactions (49). Furthermore, when a therapy abrogates novel oncogenic function primarily by disabling a gain-of-function mutation, the upper limit of the efficacy of the precision-targeted treatment will be dictated by the selection intensity that the somatic variant imparted to the cancer cell lineage. Therefore, precision treatments can be selected in clinical decision making on tumor boards to target mutations with the greatest cancer effect size, incorporating other knowledge regarding the pharmacokinetics, efficacy, side effects, and the evolution of resistance to therapies. That is, the effectiveness of a targeted therapy is expected to scale with the selection intensity of the target for any therapy that specifically inhibits the selective advantage conferred by the mutant allele. Similarly, cancer effect size can be used in the same manner to prioritize targeted development of new therapeutics, indicating the upper limit of effect for a perfect therapeutic ameliorating an oncogenic mutation. Importantly, by treating each tumorigenesis event as an independent *in situ* evolutionary trajectory, we can calculate the impact of variants within the native environment of the human soma. This inference—made directly from observations in human tissues—is complemented by experimental data using model systems

[eg, Ng et al. (50)] that are effective in providing a systematic and experimentally controlled methodology to assess functional impact of many tumorigenic genomic variants. Context-specific variation introduced via a model system causes imperfect capture of the functional impact of all variants; results from model systems should be considered alongside human cancer effect sizes.

Conclusions

The effect sizes of cancer mutations can inform nearly every aspect of basic research related to oncology, including which genes and pathways deserve greater attention in basic research. Although network interactions (51), epigenetic (52) and tumor microenvironment interactions (53), and aspects of cellular differentiation and cellular plasticity (54) mean that quantification of the selective effect of mutations does not provide an upper bound on the importance of a gene or pathway in the molecular biology of particular cases of cancer, its quantification does provide a lower bound on its importance across a cancer type, as the role of genes with a somatic variant can be no lower in importance than the selection intensity the variant imparts. By focusing on single-nucleotide variants, we have dealt with only a subset of the mutational processes observed in cancer. Importantly, effect sizes of other mutational processes, such as multinucleotide mutations, copy number alterations, or epigenetic modifications, could be similarly calculated, given sufficient prevalence data and accurate approaches to estimation of the intrinsic mutation rate for each kind of oncogenic change. Multiregion sequence datasets from large numbers of patients will surely be gathered in the future, and they promise greater power for increasingly fine-scale temporal resolutions of mutational signatures (55), rates, and fixation events (56), as well as cancer effect sizes.

Funding

This work was supported by Gilead Sciences, Inc.

Notes

Affiliations of authors: Department of Biostatistics, Yale School of Public Health, New Haven, CT (VLC, SGG, JPT); Department of Ecology and Evolutionary Biology (JPT), Yale University, New Haven, CT.

The funder had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

The authors declare no conflicts of interests.

We thank the Yale High Performance Computing Center for providing computational resources.

References

1. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495–501.
2. Martincorena I, Raine KM, Gerstung M, et al. Universal patterns of selection in cancer and somatic tissues. *Cell*. 2017;171(5):1029–1041. doi:10.1016/j.cell.2017.09.042
3. Armenia J, Wankowicz SAM, Liu D, et al. The long tail of oncogenic drivers in prostate cancer. *Nat Genet*. 2018;50(5):645–651. doi:10.1038/s41588-018-0078-z

4. Kan Z, Jaiswal BS, Stinson J, et al. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*. 2010;466(7308):869–873.
5. Zang ZJ, Cutcutache I, Poon SL, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat Genet*. 2012;44(5):570–574.
6. Kumar A, White TA, MacKenzie AP, et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc Natl Acad Sci USA*. 2011;108(41):17087–17092.
7. Zhao S, Choi M, Overton JD, et al. Landscape of somatic single-nucleotide and copy-number mutations in uterine serous carcinoma. *Proc Natl Acad Sci USA*. 2013;110(8):2916–2921.
8. Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*. 2011;471(7339):467–472.
9. Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. *Science*. 2007;318(5853):1108–1113.
10. Sjöblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. *Science*. 2006;314(5797):268–274.
11. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446(7132):153–158.
12. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214–218.
13. Evans DM, Purcell S. Power calculations in genetic studies. *Cold Spring Harb Protoc*. 2012;2012(6):664–674.
14. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*. 2007;82(4):591–605.
15. Hojat M, Xu G. A visitor's guide to effect sizes – statistical significance versus practical (clinical) importance of research findings. *Adv Health Sci Educ Theory Pract*. 2004;9(3):241–249.
16. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J*. 1986;292(6522):746–750.
17. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008;45(3):135–140.
18. Hayat MJ. Understanding statistical significance. *Nurs Res*. 2010;59(3):219–223.
19. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. *J Grad Med Educ*. 2012;4(3):279–282.
20. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting P values in the biomedical literature, 1990–2015. *JAMA*. 2016;315(11):1141–1148.
21. Leek J, McShane BB, Gelman A, Colquhoun D, Nuijten MB, Goodman SN. Five ways to fix statistics. *Nature*. 2017;551(7682):557–559.
22. Lui VWY, Hedberg ML, Li H, et al. Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer Discov*. 2013;3(7):761–769.
23. Samuels Y, Wang Z, Bardelli A, et al. High frequency of mutations of the PIK3CA gene in human cancers. *Science*. 2004;304(5670):554.
24. Nowell P. The clonal evolution of tumor cell populations. *Science*. 1976;194(4260):23–28.
25. Crespi B, Summers K. Evolutionary biology of cancer. *Trends Ecol Evol*. 2005;20(10):545–552.
26. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*. 2016;17(1):31.
27. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.
28. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375(12):1109–1112.
29. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*. 2009;25(14):1841–1842.
30. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648–660.
31. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603–607.
32. Weddington N, Stuy A, Hiratani I, Ryba T, Yokochi T, Gilbert DM. ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics*. 2008;9(1):530.
33. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317–330.
34. Cao S, Wendl MC, Wyczalkowski MA, et al. Divergent viral presentation among human tumors and adjacent normal tissues. *Sci Rep*. 2016;6(1):28294.
35. Hedberg ML, Goh G, Chiosea SI, et al. Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. *J Clin Invest*. 2016;126(1):169–180.
36. Lewis JS Jr, Beadle B, Bishop JA, et al. Human papillomavirus testing in head and neck carcinomas: guideline from the College of American Pathologists. *Arch Pathol Lab Med*. 2017;142(5):559–597. doi:10.5858/arpa.2017-0286-CP
37. Broad Institute TCGA Genome Data Analysis Center (2016): Firehose Sddata_2016_01_28 Run. Broad Institute of MIT and Harvard. http://gdac.broadinstitute.org/runs/info/DOIs_sddata.html. Accessed January 6, 2016.
38. Cannataro VL, Gaffney SG, Stender C, et al. Heterogeneity and mutation in KRAS and associated oncogenes: evaluating the potential for the evolution of resistance to targeting of KRAS G12C. *Oncogene*. 2018; 37(18):2444–2455. doi: 10.1038/s41388-017-010
39. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015; 43(Database issue):D805–D811.
40. Porta M, Crous-Bou M, Wark PA, et al. Cigarette smoking and K-ras mutations in pancreas, lung and colorectal adenocarcinomas: etiopathogenic similarities, differences and paradoxes. *Mutat Res*. 2009;682(2–3):83–93.
41. Gao J, Chang MT, Johnsen HC, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med*. 2017;9(1):4.
42. Chang MT, Asthana S, Gao SP, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol*. 2016;34(2):155–163.
43. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*. 2004;350(21):2129–2139.
44. Chapman PB, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med*. 2011;364(26):2507–2516.
45. Oyama T, Harigaya K, Sasaki N, et al. Mastermind-like 1 (MamL1) and mastermind-like 3 (MamL3) are essential for Notch signaling in vivo. *Development*. 2011;138(23):5235–5246.
46. Seiwerth TY, Zuo Z, Keck MK, et al. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clin Cancer Res*. 2015;21(3):632–641.
47. Jeffy BD, Hockings JK, Kemp MQ, et al. An estrogen receptor- α /p300 complex activates the BRCA-1 promoter at an AP-1 site that binds Jun/Fos transcription factors: repressive effects of p53 on BRCA-1 transcription. *Neoplasia*. 2005;7(9):873–882.
48. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin*. 2017; 67(1):7–30.
49. Wilkins JF, Cannataro VL, Shuch B, Townsend JP. Analysis of mutation, selection, and epistasis: an informed approach to cancer clinical trials. *Oncotarget*. 2018;9(32). doi:10.18632/oncotarget.25155
50. Ng PK-S, Li J, Jeong KJ, et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell*. 2018;33(3):450–462.e10.
51. Park S, Lehner B. Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types. *Mol Syst Biol*. 2015;11(7):824.
52. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell*. 2012;150(1):12–27.
53. DeGregori J. Connecting cancer to its causes requires incorporation of effects on tissue microenvironments. *Cancer Res*. 2017;77(22):6065–6068.
54. Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. *Nature*. 2013;501(7467):328–337.
55. de Bruin EC, McGranahan N, Mitter R, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*. 2014; 346(6206):251–256.
56. Zhao Z-M, Zhao B, Bai Y, et al. Early and multiple origins of metastatic lineages within primary tumors. *Proc Natl Acad Sci USA*. 2016;113(8):2140–2145.