Estimation of Neutral Mutation Rates and Quantification of Somatic Variant Selection Using cancereffectsizeR

Jeffrey D. Mandell[1], Vincent L. Cannataro[2], and Jeffrey P. Townsend[1,3,4,5]

ABSTRACT

Somatic nucleotide mutations can contribute to cancer cell survival, proliferation, and pathogenesis. Although research has focused on identifying which mutations are "drivers" versus "passengers," quantifying the proliferative effects of specific variants within clinically relevant contexts could reveal novel aspects of cancer biology. To enable researchers to estimate these cancer effects, we developed cancereffectsizeR, an R package that organizes somatic variant data, facilitates mutational signature analysis, calculates site-specific mutation rates, and tests models of selection. Built-in models support effect estimation from single nucleotides to genes. Users can also estimate epistatic effects between paired sets of variants, or design and test custom models. The utility of cancer effect was validated by showing in a pan-cancer dataset that somatic variants classified as likely pathogenic or pathogenic in ClinVar exhibit substantially higher effects than most other variants. Indeed, cancer effect was a better predictor of pathogenic status than variant prevalence or functional impact scores. In addition, the application of this approach toward pairwise epistasis in lung adenocarcinoma showed that driver mutations in *BRAF*, *EGFR*, or *KRAS* typically reduce selection for alterations in the other two genes. Companion reference data packages support analyses using the hg19 or hg38 human genome builds, and a reference data builder enables use with any species or custom genome build with available genomic and transcriptomic data. A reference manual, tutorial, and public source code repository are available at https://townsend-lab-yale.github.io/cancereffectsizeR. Comprehensive estimation of cancer effects of somatic mutations can provide insights into oncogenic trajectories, with implications for cancer prognosis and treatment.

Significance: An R package provides streamlined, customizable estimation of underlying nucleotide mutation rates and of the oncogenic and epistatic effects of mutations in cancer cohorts.

Introduction

Over one million distinct somatic mutations have been associated with plausible oncogenic mechanisms (1). A few hundred of these are currently clinically actionable (2), but tumor evolution frequently introduces resistance variants that overcome present therapies. The quantification of the relative effects of somatic variants has the potential to help navigate the tumor evolutionary trajectory, by informing prognosis, treatment planning, and research prioritization. However, the vast majority of somatic mutations in the genome are likely neutral (3); so much effort has been devoted to discretely identifying which of many mutations observed in tumors are "drivers" versus "passengers." Consequently, the relative strengths of intratumor-positive selection on cancer driver variants—which we term their cancer effects—have not typically been estimated.

Inferring the strength of selection from the prevalence of a tumor variant in a cancer cohort requires rigorous deconvolution of selection from baseline mutation rate across genomic sites and among patients. Baseline mutation rates can vary from tumor to tumor, depending on the impacts of previously acquired mutations, the epigenetic background, and numerous environmental factors (4–6). Analysis of gene-specific synonymous site divergence and context-specific base-pair changes from tumor-sequencing data along with tissue-specific correlates of gene mutation rates has enabled their estimation (7). Appropriate models of selection connect these baseline mutation rates with observed variant prevalence in tumor cohorts, distinguishing highly oncogenic variants from variants that occur frequently within cancer cells due to high underlying rates of mutation (3). Accurate estimation of the strength of selection on driver mutations based on explicit selective models is essential to the advancement of tumor sequence data analysis (8).

To facilitate this deconvolution of prevalence into baseline mutation rate and scaled selection coefficient, we have developed cancereffectsizeR, an R package that organizes somatic variant data, calculates sample- and site-specific mutation rates, and quantifies cancer effects under customizable models of selection. The cancereffectsizeR package provides an extensible framework to refine our understanding of somatic mutations beyond the traditional dichotomy of selected drivers and neutral passengers. Recent authors have proposed additional discrete categories such as "superdrivers," weak drivers, and impactful passengers (9, 10); cancereffectsizeR takes the next step, enabling inference of a continuous range of effects that may vary depending on epistatic interactions, tumor grade and subtype, or other clinically relevant factors. In this report, we describe core package features and their methodology, and we emphasize the broad support supplied by the package for user-provided data and highly customized analyses of variant selection. We also demonstrate that inferences of cancer effect are robust to method of mutation rate calculation and that known cancer-associated variants tend to be of higher effect than other variants—even when they appear in cancer cohorts at lower prevalence.

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut. [2]Department of Biology, Emmanuel College, Boston, Massachusetts. [3]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut. [4]Genetics, Genomics, and Epigenetics Research Program, Yale Cancer Center, New Haven, Connecticut. [5]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut.

Corresponding Author: Jeffrey P. Townsend, Yale School of Public Health, 135 College Street, Suite 200 #222, New Haven, CT 06525. E-mail: jeffrey.townsend@yale.edu

Cancer Res 2023;83:500–5

doi: 10.1158/0008-5472.CAN-22-1508

Materials and Methods

Loading mutation data

By integrating somatic variant data from whole-genome, whole-exome, and targeted sequencing experiments, cancereffectsizeR obtains especially high power to estimate cancer effects. In cancereffectsizeR, the estimation of mutation rates and cancer effects begins with loading high-confidence somatic variant data in Mutation Annotation Format (MAF; **Fig. 1**). A "preloading" feature checks MAF files for common problems, flags possible false-positive records, and converts variant coordinates between genome builds via the rtracklayer (RRID:SCR_021325) package's liftOver (RRID:SCR_018160) functionality. Another function identifies pairs of samples with suspiciously high variant overlap, which can help to detect inadvertent sample duplicates in analyses that use multiple sources of sequencing data.

Users can define which genomic regions are covered by each data source so that effect-size inferences correctly account for which variants can possibly be found in each sample group. Once data are loaded, variants can be viewed with gene and transcript annotations, filtered, or combined into "compound variants" that are treated as single entities for mutation rate calculation and selection inference. Clinical or epidemiological data can be loaded into the analysis to allow further steps in the workflow to be run on arbitrary groups of samples with tailored parameters.

Mutation rates and selection

Mutation rates are computed by convolving gene-by-gene estimates of mutation rates from synonymous site rates and covariates with trinucleotide site-by-site rates based on mutational signature extraction. Gene-by-gene mutation rates make use of the dNdScv regression model (11) with customizable mutation rate covariates. We have assembled sets of covariates for each of 20 tumor types that include gene expression, chromatin marks, and replication timing data. We also provide a simple workflow to generate custom covariates. Noncoding regions are included by combining trinucleotide rates associated with noncoding sites with the locus-specific mutation rate of the nearest gene. Alternatively, users may input their own precalculated gene mutation rates.

Mutational signature extraction is used to attribute observed tumor variants to a linear combination of mutational signatures, such as the subset of COSMIC signatures (1) that are relevant to the tissue type. Extraction is conducted with MutationalPatterns (the default; ref. 12) or deconstructSigs (13), or users may input precalculated signature weights from other sources. Some signatures have been attributed artifactual status as consequences of sample preparation rather than biological processes (14); as previously described (15), cancereffectsizeR corrects for these artifactual signatures, re-normalizing non-artifactual signature weights to yield relative rates of substitution in all trinucleotide contexts. In a given tumor and gene, the gene mutation rate is partitioned across all sites in the gene in accordance with these context-informed relative rates, yielding tumor-specific rates for all possible substitutions (7). Targeted-sequencing data are assigned gene mutation rates from specified exome and/or genome-sequenced data, and trinucleotide mutation rates are determined from a group averaging of signatures found in the companion data.

Variant-specific scaled selection can be estimated under several models of selection. The default model assumes constant selective pressure on all mutations over oncogenesis, and no epistatic effects. In addition, a pairwise epistatic model enables inference of selection with epistatic effects between pairs of variants (bioRxiv 2022.01.20.4771322022). Models of greater complexity, such as those incorporating selective epistasis, require larger sample sizes of tumor sequence than the default model. Models can be applied to single-nucleotide variants (SNV) such as noncoding substitutions and amino-acid–changing substitutions, or they can be applied to ensembles of nucleotide variants that are all assumed to cause the same cancer effect—such as within functional domains or genes. The effect-size calculation step also permits input of custom models of selection, including models that feature alternate kinds of user-supplied data.

Reference data

cancereffectsizeR includes functions to generate custom reference data for almost any genome build or species for which genome and transcript definitions can be supplied in common file formats. We include straightforward instructions for specifying custom mutational signature definitions and generating mutation rate covariates from user-supplied molecular data. We have also provided companion data packages, *ces.refset.hg38* and *ces.refset.hg19*, that include gene and coding sequence definitions compatible with the hg38/hg19 human genome builds, as well as COSMIC mutational signature definitions and precomputed mutation rate covariates for twenty tissue types. Additional package functions, detailed in the online

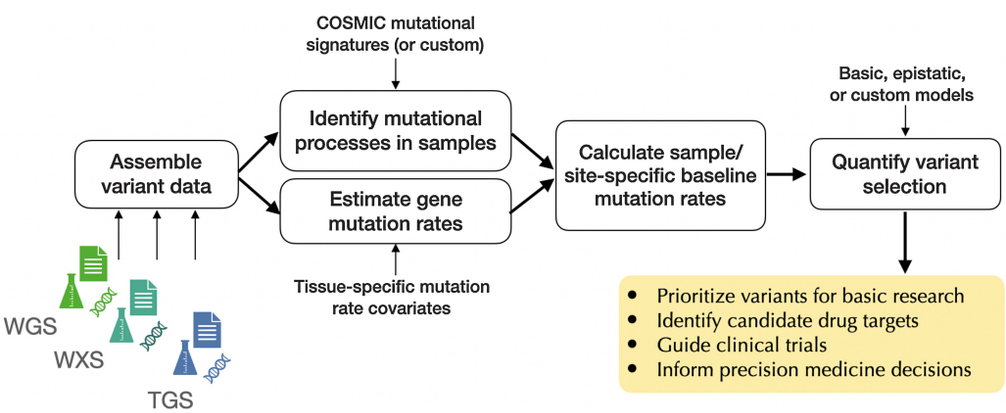


Figure 1.
The cancereffectsizeR workflow, spanning assembly of diverse variant datasets to quantification of effect sizes.

reference manual, support the use of cancereffectsizeR with tissue types for which mutational signature definitions or mutation rate covariates are unavailable.

Data availability

The data analyzed in this study were obtained from The Cancer Genome Atlas (TCGA; https://portal.gdc.cancer.gov; data release 34.0) and cBioPortal (ID: luad_mskcc_2020).

Results

To illustrate a cancereffectsizeR workflow, we loaded somatic mutation calls from the TCGA LUAD (lung adenocarcinoma) project and calculated gene mutation rates using lung tissue covariates from the companion reference data package. For the mutational signature extraction step, we used COSMIC 3.2 mutational signature definitions, with signatures presumed absent in lung tissue excluded. We excluded these signatures to reduce "bleeding" between signatures with similar profiles and to help ensure biological plausibility of fitted signatures as recommended by Alexandrov and colleagues (14). We applied the default model of selection to all recurrent variants, and we found that 19 of the 20 highest-effect variants were amino-acid–changing substitutions in the three oncogenes *BRAF, KRAS*, and *EGFR* and the tumor-suppressor *TP53* (**Fig. 2A**). Out of these top variants, prevalence of the 7 variants in *BRAF, KRAS*, and *EGFR* ranged from 3 to 54 (median 17), and from 2 to 6 (median 3) in the 12 *TP53* variants. That variants in these well-known cancer drivers received the highest effect estimates, yet span a broad range of prevalences, is an indication of the successful deconvolution of mutation rate and selection.

Next, we tested for pairwise epistasis between *KRAS*, *EGFR*, *BRAF*, and *TP53*. Because analyses of epistasis require larger datasets for sufficient power to estimate more parameters, we augmented the TCGA LUAD data with targeted sequencing of 604 lung adenocarcinoma tumors available from cBioPortal (ID: luad_mskcc_2020). We applied a pairwise epistatic model (bioRxiv 2022.01.20.477132) enforcing shared scaled selection coefficients across all observed nonsynonymous and splice-site mutations within each gene. Analyzed under a model of selective epistasis, selection for mutations in each gene depended on the mutation status of driver sites in the other gene. To isolate pairwise epistasis—that is, to avoid higher-order selective interactions between *KRAS*, *EGFR*, *BRAF*, and *TP53*—each pairwise

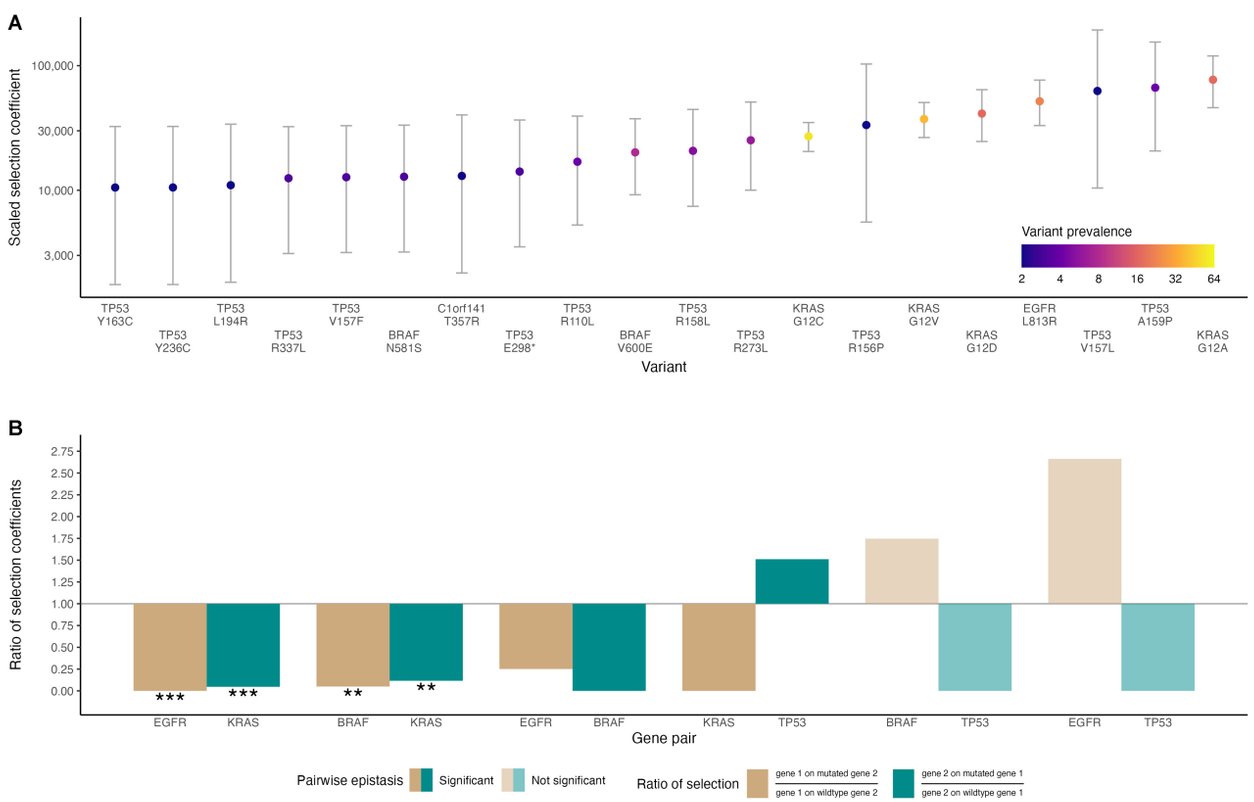


Figure 2.

Selection inferences from a standard cancereffectsizeR workflow (version 2.6.5) with somatic variant data from exome and panel sequencing of lung adenocarcinoma. **A,** Highest effect recurrent somatic variants (and 95% confidence intervals) under the default model of selection at individual genomic sites. **B,** Ratios of selection coefficients for the observed nonsynonymous and splice-site mutations in gene one after mutation of gene two relative to selection coefficients of gene one when other genes analyzed are unmutated (tan bars), and ratios of selection coefficients for the observed nonsynonymous and splice-site mutations in gene two after mutation of gene one relative to selection coefficients of gene two when other genes analyzed are unmutated (green bars). For some gene pairs, the epistatic model is not significantly better than a model that assumes no epistatic effects ($P > 0.05$, likelihood ratio test; transparent bars). Asterisks denote genes within pairs that not only are inferred to be subject to selective pairwise epistasis, but that also exhibit specific statistically significant directional changes in selection after mutation in the other gene. **, $P < 0.01$; ***, $P < 0.001$; likelihood ratio test.

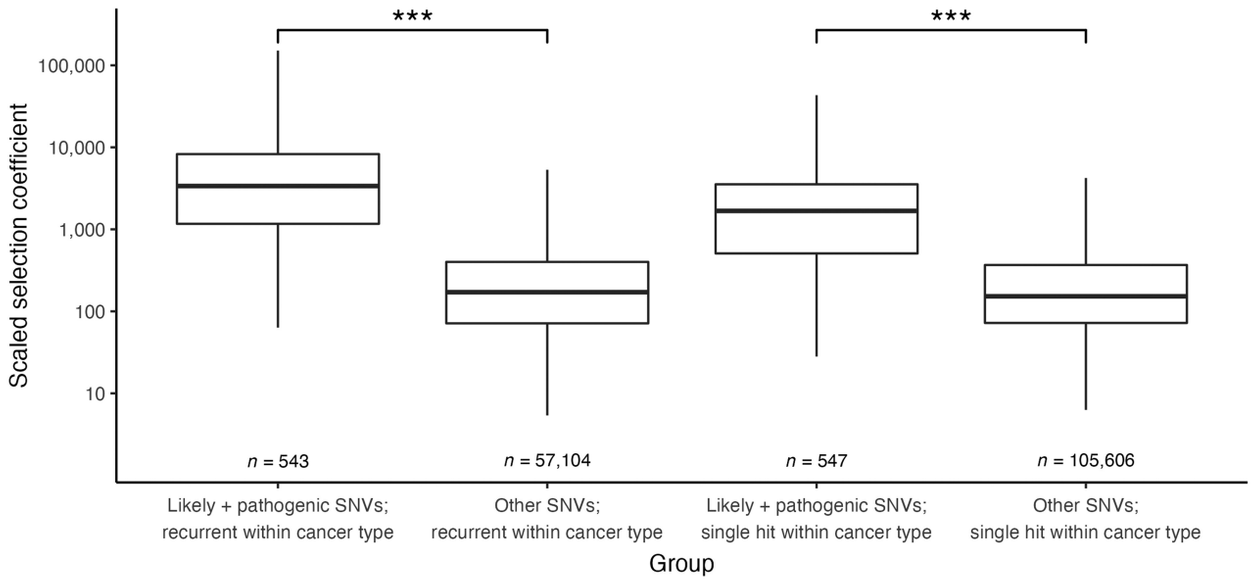


Figure 3.

Boxplots of the cancer effects of variants appearing in two or more patients across eight TCGA cohorts. A set of merged somatic variants that are annotated within ClinVar as likely pathogenic or pathogenic is compared with other variants, and sites mutated recurrently within a cancer type are compared with sites hit only once within a cancer type. Each cancer effect estimate is a cancer type–specific inference; variants appearing in multiple cohorts are reported by multiple estimates. Two statistically significant pairwise comparisons are shown, but all possible pairwise comparisons of groups yielded statistically significant differences (Mann–Whitney *U* test, $P < 10^{-16}$ for all). ***, $P < 0.001$.

inference in this analysis only included samples that lacked mutations in whichever of the four genes were not the subject of the inference. Driver mutations in pairs of RAS pathway genes (*KRAS, EGFR*, or *BRAF*) exhibited substantial antagonistic epistasis ($P < 0.01$ for each pair, likelihood ratio test; **Fig. 2B**). In *BRAF/KRAS* and *EGFR/KRAS*, selection for alterations in each gene was substantially reduced after mutation in the other ($P < 0.01$ for all four, likelihood ratio test). The best estimate for the *EGFR/BRAF* is similar, and antagonistic epistasis between the pair is statistically significant. The significant epistatic effects suggest reduced selection in one or both genes, but without statistical significance to the specific reduction of selection in either gene individually. This antagonistic epistasis found among these three oncogenes is consistent with their common roles activating the MAPK/ERK pathway. Epistatic effects did not achieve significance for gene pairs involving *TP53* except for *TP53/KRAS* ($P < 0.001$, likelihood ratio test). Best estimates of selective epistatic effects between these genes and the tumor-suppressor *TP53* exhibit a different pattern that suggests sensitivity of the evolutionary trajectory to mutation order.

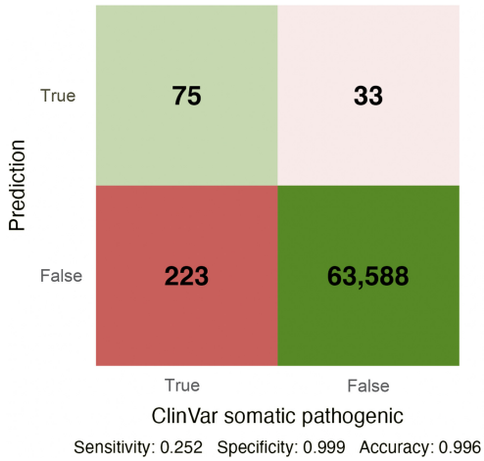
To validate that cancer effect estimates provide novel and useful information about cancer relevance, we compared effect sizes for variants with ClinVar annotations of somatic pathogenic or likely pathogenic to other variants in eight TCGA cancer cohorts (endometrial carcinoma, UCEC; colon adenocarcinoma, COAD; lung squamous-cell carcinoma, LUSC; LUAD; breast carcinoma, BRCA; skin cutaneous melanoma, SKCM; bladder urothelial carcinoma, BLCA; head-and-neck squamous-cell carcinoma, HNSC). We estimated cancer type–specific effects for all SNVs that appeared in at least two patients in the pan-cancer dataset. Cancer effects tended to be substantially higher in likely pathogenic or pathogenic SNVs ($P < 2.2 \times 10^{-16}$; Mann–Whitney *U* test): The median likely pathogenic or pathogenic SNV effect ranked at the 97.9th percentile of all effect estimates. Strikingly, likely pathogenic or pathogenic SNVs that were single hits within their cancer cohorts were generally much higher ranked by cancer effect than were other SNVs that were recurrent, with median effects at the 96.5th and 52.0th percentiles, respectively (**Fig. 3**).

The ClinVar designation of somatic pathogenic variants reflects current knowledge. Therefore, ClinVar-designated variants are an as-yet incomplete set of cancer-related variants. One would expect that more prevalent variants are more likely to have been discovered and included. To verify that cancer effects are powerful predictors of known cancer association, we multiply regressed likely pathogenic or pathogenic ClinVar status against mean prevalence (across the 8 TCGA projects), highest prevalence (within a cancer cohort), mean cancer effect (across cancer types, assigning an effect of 0 for cancer types without the variant), highest cancer effect, and also protein function impact scores from SIFT and PolyPhen-2 (**Fig. 4**). A bootstrapped dominance analysis (16) found, consistently across all 100 bootstraps, that the cancer effect predictors had general dominance over all other predictors; that is, they had the highest average contributions to prediction across all sizes of submodels.

Cancer effect estimates are robust to reasonable changes in mutation rate calculation methods. We compared TCGA LUAD effects as previously estimated with estimates produced using two other workflows: First, using gene mutation rates calculated with MutSigCV instead of dNdScv, and second, with deconstructSigs instead of MutationalPatterns for signature analysis. Effect estimates were highly correlated with the original estimates (Pearson's $r = 0.87$ and 0.99, respectively; Supplementary Fig. S1).

Discussion

As more tumor-sequence data are collected, increasingly precise quantification of variant effects is possible, estimated with ever-more



Parameter	Estimate	SD	*P*
(Intercept)	−7.46	0.149	<0.001
Mean cancer effect	3.30	0.434	<0.001
Top cancer effect	−1.67	0.419	<0.001
SIFT score	−0.55	0.197	0.005
Polyphen-2 score	0.50	0.106	<0.001
Mean prevalence	0.26	0.134	0.053
Top prevalence	0.19	0.126	0.143

Predictor importance:
Mean effect > Top effect > Mean prevalence > Top prevalence > PolyPhen-2 > SIFT

Figure 4.
Contingency table (confusion matrix) and model summary of a multiple logistic regression predicting merged pathogenic or likely pathogenic ClinVar status of variants based on mean cancer effect across eight cancer types, top cancer effect across eight cancer types, SIFT score, PolyPhen-2 score, mean prevalence across eight cancer types, and top prevalence across eight cancer types. Noncoding variants—which lacked SIFT and PolyPhen-2 scores—were excluded from the regression. Cancer effect measures were log transformed, and all predictive parameters were standardized. Predictor importance was determined from bootstrapped dominance analysis (100 bootstrap runs); each predictor exhibited pairwise general dominance over all less important predictors.

informative models of oncogenesis. Cancer effects can be estimated for specific cohorts of tumors; the effect estimate averages over known or unknown heterogeneities within the designated cohort, such as differing tumor microenvironments. Epistatic models can evaluate how selection in each of a pair of variants or genes depends on the status of the other, explaining observations of mutual exclusivity, co-mutation, and variation in disease trajectories. Support for arbitrary sample grouping—for example, by chromosomal variant or copy-number status—enables hypothesis-driven analysis of differential substitution effects. In contrast with the dichotomous classification of drivers and passengers, the continuous scale of cancer effect provides a clear, principled prioritization of drivers: Within a cancer cohort, the variants with the highest effects have the greatest per-patient contribution to aberrant cellular proliferation.

As an analytic framework, cancereffectsizeR is constructed to facilitate continued development to broaden the precision and scope of inference of cancer effect sizes, and to meet evolving user needs. For example, stage/grade-specific or pre/post-treatment models could demonstrate which variants play key roles in serial phases of disease—including in the evolution of therapeutic resistance. In addition, indicators of altered mutability—such as genome-wide hypermutation signatures (17) or regional chromatin structure (18)—can be included to increase the precision of the sample-specific mutation rate calculation. Incorporation of these effects could be especially helpful in cancer types with highly variable mutational burden. Application of phylogenetic approaches would provide orthogonal information regarding mutation order, which could constrain and therefore further improve the precision of inference regarding epistatic selection. The search space of possible epistatic relationships can be examined by large-scale data analysis; alternatively, a two-phase analysis can use principled approaches to identify likely sets of cooperating variants (19) and then quantify their selective epistasis.

Single-base substitutions are the most prevalent small variants. Future work is intended to extend the package, enabling estimation of the cancer effects of doublet-base substitutions and small insertions/deletions. Other variant types that might be crucial to oncogenesis in many types of cancer, such as copy-number alterations, loss of heterozygosity, and epigenetic phenomena present greater challenges to determination of the rate at which they occur in cancer-competent cells. Future extensions of cancer effect estimation to a comprehensive range of somatic genetic alterations will enable thorough accounting of oncogenic trajectories, with implications for prognosis, prioritization of research, drug targeting, and treatment planning.

Authors' Disclosures

V.L. Cannataro reports personal fees from Black Diamond Therapeutics outside the submitted work. J.P. Townsend reports personal fees from Agios Pharmaceuticals and Black Diamond Therapeutics outside the submitted work. No disclosures were reported by the other author.

Authors' Contributions

J.D. Mandell: Conceptualization, software, formal analysis, visualization, methodology, writing–original draft. **V.L. Cannataro:** Conceptualization, software, methodology, writing–review and editing. **J.P. Townsend:** Conceptualization, formal analysis, supervision, methodology, project administration, writing–review and editing.

Acknowledgments

The funding to support this research was provided (to J.P. Townsend) by NIH 1P50DE030707 and the Elihu Professorship endowment.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Note

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Received May 10, 2022; revised October 11, 2022; accepted November 30, 2022; published first December 5, 2022.

References

1. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47:D941–7.
2. Prawira A, Pugh TJ, Stockley TL, Siu LL. Data resources for the identification and interpretation of actionable mutations by clinicians. Ann Oncol 2017;28: 946–57.
3. Cannataro VL, Townsend JP. Neutral theory and the somatic evolution of cancer. Mol Biol Evol 2018;35:1308–15.
4. Starrett JH, Guernet AA, Cuomo ME, Poels KE, van Alderwerelt van Rosenburgh IK, Nagelberg A, et al. Drug sensitivity and allele specificity of first-line osimertinib resistance mutations. Cancer Res 2020;80:2017–30.
5. Schuh A, Becq J, Humphray S, Alexa A, Burns A, Clifford R, et al. Monitoring chronic lymphocytic leukemia progression by whole-genome sequencing reveals heterogeneous clonal evolution patterns. Blood 2012;120:4191–6.
6. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, et al. The DNA methylation landscape of advanced prostate cancer. Nat Genet 2020;52:778–89.
7. Cannataro VL, Gaffney SG, Townsend JP. Effect sizes of somatic mutations in cancer. J Natl Cancer Inst 2018;110:1171–7.
8. Bozic I. Quantification of the selective advantage of driver mutations is dependent on the underlying model and stage of tumor evolution. Cancer Res 2022;82: 21–4.
9. Grossmann P, Cristea S, Beerenwinkel N. Clonal evolution driven by superdriver mutations. BMC Evol Biol 2020;20:89.
10. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, et al. Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. Cell 2020;180:915–27.
11. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. Cell 2017;171: 1029–41.e21. https://doi.org/10.1016/j.cell.2017.09.042.
12. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med 2018;10:33.
13. 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;17:31.
14. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578: 94–101.
15. Cannataro VL, Mandell JD, Townsend JP. Attribution of cancer origins to endogenous, exogenous, and preventable mutational processes. Mol Biol Evol. Oxford Academic; 2022;39:msac084.
16. Azen R, Budescu DV. The dominance analysis approach for comparing predictors in multiple regression. Psychol Methods 2003;8:129–48.
17. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell 2012; 149:979–93.
18. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature 2020;578: 112–21.
19. Klein MI, Cannataro VL, Townsend JP, Newman S, Stern DF, Zhao H. Identifying modules of cooperating cancer drivers. Mol Syst Biol 2021;17:e9810.