Testing of custom evolutionary models of oncogenesis

Extension to double-base substitutions

Software

Reproducible analysis and data

Jeffrey D. Mandell[1] and Jeffrey P. Townsend[1,2,3]

1) Program in Computational Biology and Bioinformatics, Yale University
2) Department of Biostatistics, Yale School of Public Health
3) Department of Ecology and Evolutionary Biology, Yale University

Contact: Jeff.Mandell@yale.edu
Acknowledgment: We thank the JMR Barker Foundation for funding enabling performance of this research.

Summary

We previously developed the cancereffectsizeR software package to quantify somatic selection for single-base substitution (SBS) mutations in tumor sequencing data sets, with extensive support for customized analyses[1,2]. In cancer, strength of selection—or "cancer effect"—indicates the degree to which variants contribute to aberrant growth and proliferation. Here, we present methods that extend this quantification of cancer effects to double-base substitutions (DBS). In an analysis of The Cancer Genome Atlas's (TCGA) skin cutaneous melanoma (SKCM) cohort, we validate that estimated DBS effects are in line with estimated effects of SBS variants that induce equivalent changes to protein-coding sequence.

Background

Cancer tissues are often characterized by distinctive somatic alterations to their genomes, from single base substitutions to large-scale structural variants. As stated in a broadly accepted automative metaphor, most somatic mutations are neutral passengers, while a subset of mutations act as drivers that confer phenotypes associated with disease progression. Mutations that are observed to be frequent in cancer cohorts are natural candidate drivers. However, variation in cellular mutation rates across genomic sites—and among patients with varying mutagenic exposure—can obscure the relative importance of different variants. Estimation of patient– and site–specific mutation rates enables an assessment of the strength of selection on variants contributing to oncogenesis under customizable models of somatic evolution[2], enabling research prioritization and treatment planning. In a report on the cancereffectsizeR software[2], we showed that variants of known cancer relevance are substantially more strongly selected than most other variants (Fig. 1), and that strength of selection is a better predictor of a variant's known cancer relevance than cohort prevalence or protein function impact scores (SIFT, PolyPhen-2). Since DBS occur at different rates than SBS and produce different changes in coding sequences, no approach for quantifying somatic selection for DBS has yet been implemented.

Methods

- Developed cancereffectsizeR functionality to annotate DBS in somatic variant call sets with amino acid consequence and other information needed to support integrated analysis of DBS and SBS.
- Adapted negative binomial regression framework from Martincorena et al.[3]—which was built to estimate gene-specific rates of insertion/deletion (indel) occurrence—to estimate neutral DBS rates by gene.
- Accounted for variability in DBS acquisition by sequence context (e.g., AC→GT) and combined with gene-specific DBS rates to yield rate estimates for any DBS across the exome.
- Incorporated DBS rates into a model of clonal selection that allows inference of a shared scaled selection coefficient via maximum likelihood estimation for observed DBS variants in tumor sequencing cohorts.
- Applied the framework to estimate DBS effects in TCGA SKCM, as SKCM has a high DBS burden. Some amino-acid changes (e.g., BRAF V600E) were induced by both SBS and DBS in this cohort. Therefore, we compared selection inferences for these SBS and DBS, with the expectation of strong correlation.
- Compared the intraclass correlation of effect estimates for cognate SBS versus DBS (same codon change); SBS versus SBS (same codon change); and SBS versus SBS (same codon, synonymous).

Results

- Of 9,623 SKCM DBS (observed in 465 patients), 171 were identified at sites that had equivalent SBS amino-acid changes in the cohort. None of these DBS were synonymous.
- There were 246 codons that were affected by two distinct SBS with equivalent protein consequences (prevalences of each SBS in the cohort varied). 112 of the codon changes were non-synonymous, and 134 were synonymous.
- SBS and DBS effect estimates at matched sites exhibited a Pearson's correlation of 0.79 (Fig. 2).
- DBS and SBS effects exhibited higher intraclass correlation than did SBS pairs (Table 1).

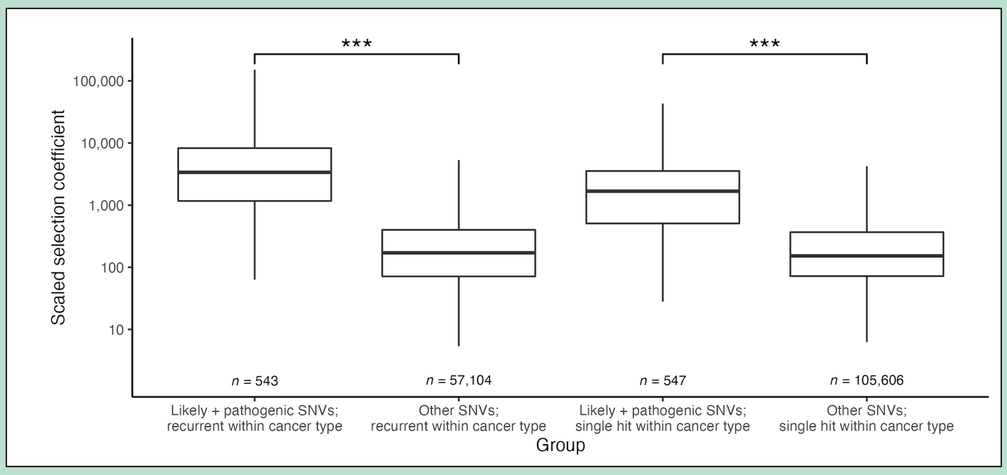

Figure 1

Box plots of the cancer effects of variants appearing in two or more patients across eight TCGA cohorts, separated by ClinVar somatic pathogenic annotation and recurrence within cohort. Estimates are specific to cancer type; variants appearing in multiple cohorts appear as multiple estimates. Reprinted from Mandell, Cannataro, and Townsend[2].

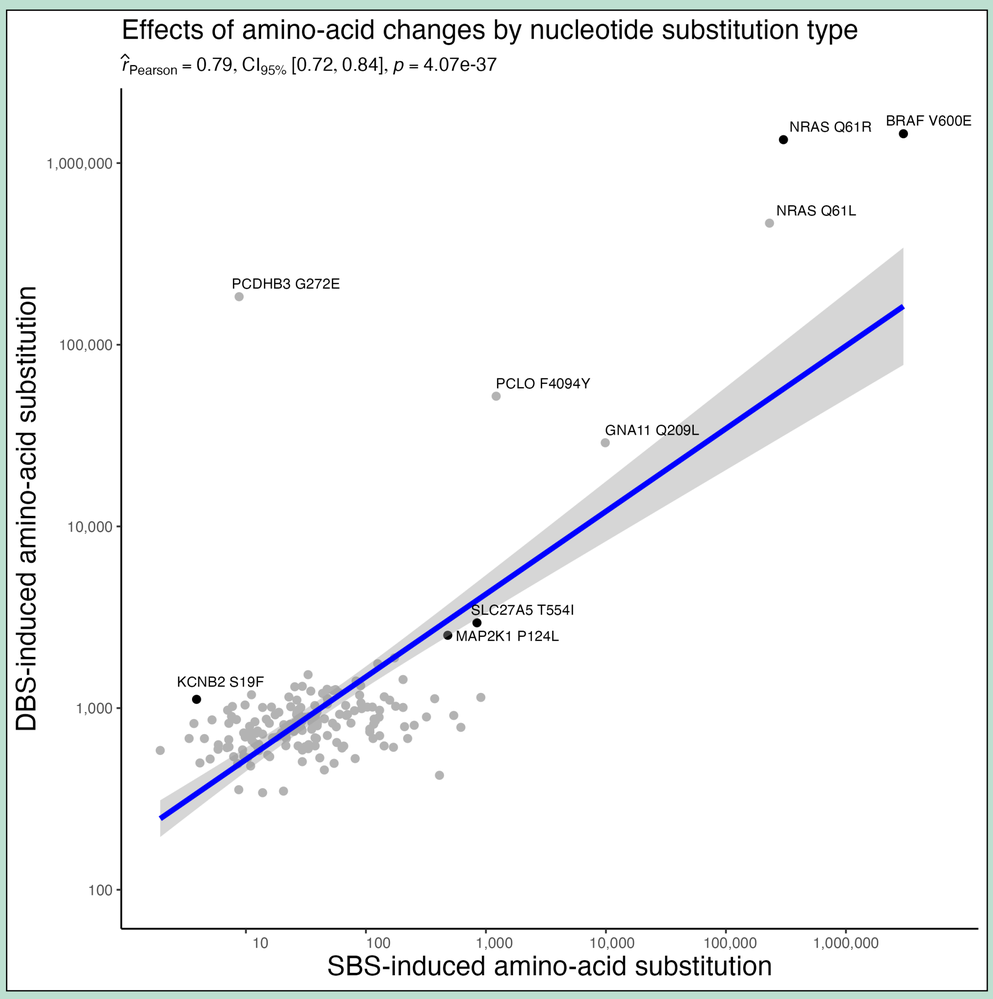

Figure 2

Comparison of cancer effects (scaled selection coefficients) for SBS and DBS variants that produced equivalent amino-acid changes, some of which were induced by DBS in multiple SKCM patients (marked black).

Discussion

DBS effect estimates tracked well with effect estimates of equivalent coding SBS, even though most of the analyzed DBS occurred in single patients. Estimates of selection for same-codon-change SBS are less accurate than estimates at the amino acid substitution level (which would incorporate both SBS in the inference), which may explain the moderate correlation of these paired effects. The lack of correlation between effect estimates of same-codon synonymous SBS may reflect that most of these variants are probably neutral (hence selection is randomly overestimated), while any variants under selection may act through distinct non-protein-coding mechanisms.

DBS effect inference could be improved by employing a mutational signature approach—such as extraction of COSMIC DBS signatures—to determine sample-specific relative rates for various dinucleotide substitutions (e.g., reflective of differential UV exposure). This approach may require WGS data, as exomes may have insufficient DBS burden. Another improvement would be to incorporate synonymous DBS into the gene rate model as markers of neutrality, akin to SBS dN/dS. DBS gene rates may be overestimated in this analysis, as the model assumes that DBS outside of purported "cancer genes" are neutral.

Further validations that DBS effect is a measure of cancer relevance should be conceived and tested. A thorough analysis of the DBS selective landscape in melanoma and other cancer types is underway.

Comparison	Number of sites	Intraclass correlation	95% CI
DBS vs. SBS, same codon change	171	0.73	(0.65, 0.79)
SBS vs. SBS, same codon change	112	0.41	(0.24, 0.55)
SBS vs. SBS, same codon synonymous	134	-0.05	(-0.21, 0.12)

Table 1

Comparison of scaled selection coefficients estimated at genomic sites where equivalent amino-acid changes occur in SKCM via an SBS and DBS mutation, or a pair of SBS mutations.

References

1. Cannataro, V.L., Gaffney, S.G., Townsend, J.P., 2018. Effect Sizes of Somatic Mutations in Cancer. JNCI: Journal of the National Cancer Institute 110, 1171–1177. https://doi.org/10.1093/jnci/djy168
2. Mandell, J.D., Cannataro, V.L., Townsend, J.P., 2023. Estimation of Neutral Mutation Rates and Quantification of Somatic Variant Selection Using cancereffectsizeR. Cancer Research 83, 500–505. https://doi.org/10.1158/0008-5472.CAN-22-1508
3. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., Campbell, P.J., 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042