VIEWPOINTS

Wagging the long tail of drivers of prostate cancer

Vincent L. Cannataro[1], **Jeffrey P. Townsend**[1,2,3]*

1 Department of Biostatistics, Yale University, New Haven, Connecticut, United States of America, **2** Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **3** Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, United States of America

* Jeffrey.Townsend@Yale.edu

Citation: Cannataro VL, Townsend JP (2019) Wagging the long tail of drivers of prostate cancer. PLoS Genet 15(1): e1007820. https://doi.org/10.1371/journal.pgen.1007820

Editor: Matthew L. Freedman, Dana Farber Cancer Institute, UNITED STATES

Published: January 17, 2019

Funding: Funding to support this research was provided by the Notsew Orm Sands Foundation and NIH R01 LM012487. The funders had no role in the preparation of the article.

Competing interests: The authors have declared that no competing interests exist.

Armenia and colleagues [1] recently analyzed the largest set of prostate cancer exomes to date—1,013 exomes from 680 primary and 333 metastatic tumors. Whereas it has been suggested that gene discovery is near saturation for localized nonindolent prostate cancer [2], Armenia and colleagues demonstrated the utility of uniformly analyzing data sets from a large number of clinically diverse tumors from the same tissue. Through their analysis, they revealed 97 significantly mutated genes and found that the prevalences of these mutated genes follows a long tail, with a few genes containing a substitution in comparatively many tumors, and many genes containing a substitution in few tumors. This "long tail" distribution of significantly mutated genes suggests that increases in sample size still lead to the discovery of rarely mutated but significant drivers. Indeed, the long-tail distribution of potential drivers gives hope in our fight against cancer, increasing the number of genes and pathways that might be productively therapeutically targeted by precision medicine.

The imagery of the long tail elicits not just the idea that in each cancer type there continue to be genes that affect tumorigenesis and cancer development that can still be discovered by continued sequencing. It also inherently evokes the relative ranking of those genes and their importance in explaining cancer. Armenia and colleagues follow a well-developed decade-long tradition of analysis of tumor sequence data to identify drivers—a tradition that has used two ways to describe this long tail. One way, exemplified by Lawrence and colleagues [3], ranks members of the tail of cancer genes discovered by their *P* value. However, if one's goal is to uncover and propose the relative contribution of genes implicated in cancer to the cancer phenotype (i.e., genetic alterations contributing to increased cellular division or cellular lifetime), then *P* values are not an appropriate metric, because *P* values are thresholds of belief and not measures of effect [4].

Armenia and colleagues follow a second approach that is also common: ranking genes discovered by the prevalence of the mutations that are observed at high frequencies (because of the number of cells in a typical cancer, all mutations observed at statistically significant allele frequencies via high-throughput sequencing are present within a large number of cells within the cancer; otherwise they would be exceedingly unlikely to be observed even with deep sequencing). Ranking genes by prevalence makes more sense than ranking by *P* value. Given statistical significance, genes that are most prevalent are those that affect tumorigenesis and cancer development in the greatest number of patients. Therefore the relative prevalence of mutations observed at high frequencies within tumors (the long tail reported by Armenia and colleagues) usefully conveys how many patients might benefit from a therapeutic that targets the mutant state of the gene.

There is now, however, a third way to rank drivers of a cancer type based on mutations observed at high frequencies in tumor sequence data, a way that perhaps has been in the back of the mind of scientists all along as they have thought about cancer drivers. They can now be ranked by their relative contribution to survival and proliferation of the cancer; i.e., the intensity by which the mutated lineages are naturally selected to expand and reach detectable levels in the neoplasm and/or tumor. This measure compares the actual flux of substitutions observed to the expected flux of substitutions in the absence of selection and is the equivalent of the "scaled selection coefficient" or "selection intensity" used in population genetic research [5–7]. Selection intensity conveys how much a patient might benefit from a targeted therapeutic that fully abrogated the gain-of-function associated with an oncogenic mutation.

Further analysis of the extensive data gathered by Armenia and colleagues on prostate cancer, estimating selection intensity on specific single nucleotide mutations [5,6], and ranking observed mutations by selection intensity yields a long tail of cancer driver mutations with a very similar overall shape to that seen when ranking genes by prevalence as in Armenia and colleagues (Fig 1; S1 Table). However, ranking observed mutations by selection intensity provides a markedly discrepant ranking of the drivers. This discrepancy in rank is the consequence of enormous variation of mutation rate (among both genes and trinucleotides) and mutational target size (tumor suppressors usually can be disabled in many ways, but proto-oncogenes require very specific mutations to become oncogenes). Genes such as *AR* and *KMT2C*, with comparatively high mutation rates and many different oncogenic mutations present within tumor samples, have a lower selection intensity for specific single nucleotide variants and therefore a lower comparative ranking versus a ranking by the total number of substitutions within the gene among tumors samples. Other genes, such as *CUL3* and *BRAF*, have relatively low mutation rates and feature recurrent substitutions of one specific amino acid change, and therefore the selection for these variants must be comparatively high in order for us to observe these mutations at detectable levels within tissue from numerous tumors.

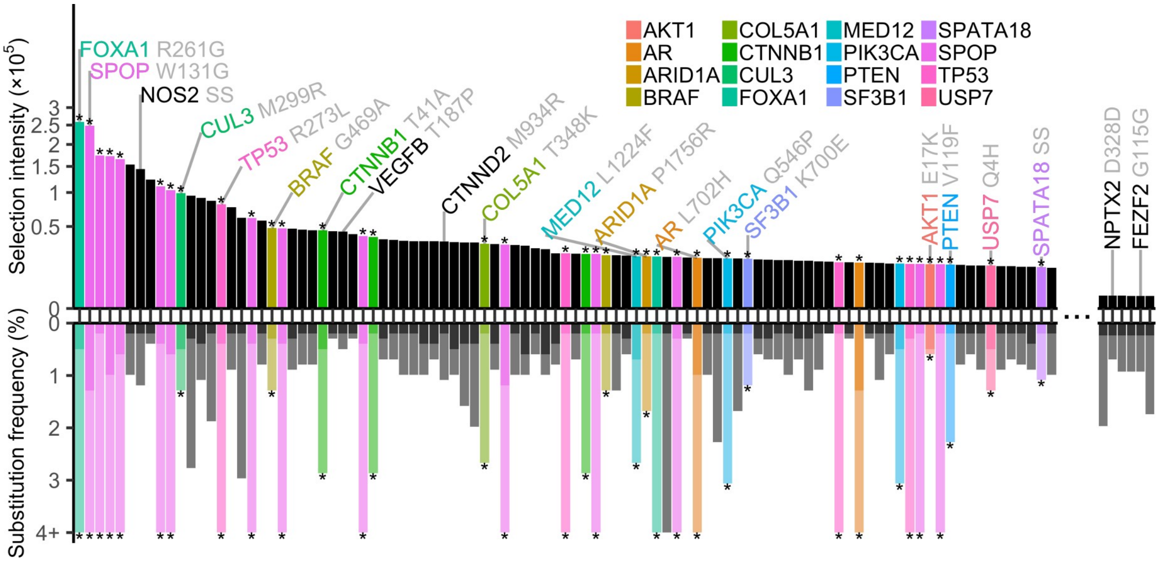

Fig 1. Selection intensity and substitution frequency of the 97 recurrent single nucleotide variants out of the 890 available in the Armenia and colleagues' data set with the highest selection intensity and the 6 with the lowest selection intensity (all 6 are synonymous substitutions). Bars above the *x*-axis convey the selection intensity (plotted here on a square-root scale), with significantly mutated genes (as designated by Armenia and colleagues) labeled with an asterisk (*) and plotted each with a unique color. Bars below the *x*-axis convey the prevalence of the mutation at the gene level (semitransparent), overlaid with the prevalence of the specific single nucleotide variant ranked (solid). The NOS2 SS label refers to a substitution at chromosome 17, nucleotide position 26087772, and the SPATA18 SS label refers to a substitution at chromosome 4, nucleotide position 52946086 (hg19 coordinates). SS, splice site.

https://doi.org/10.1371/journal.pgen.1007820.g001

Armenia and colleagues are prescient about the importance of the novel cancer drivers that they discuss, which appear in their long tail at low prevalence but which turn out to be genes that have high impact on tumorigenesis and cancer development. Recurrently substituted sites in *CTNNB1*, *CUL3*, *BRAF*, *ARID1A*, and *SF3B1* described within Armenia and colleagues are within the top selection intensities calculated (Fig 1). Several recurrently mutated single nucleotide variants with previously described associations with prostate cancer and/or metastatic phenotypes, such as *NOS2* [8], *CTNND2* [9], and *VEGFB* [10], were not found to be significantly mutated at the gene level yet nevertheless are estimated to have remarkably high selection intensities, illustrating that there is much yet to learn about the potential translationally relevant genetic basis of some prostate cancers with additional tumor sequencing and even larger sample sizes.

How is it that the genes that Armenia and colleagues picked out to discuss, among many from far down their long tail in prevalence, would turn out to be genes of high selection intensity? Presumably, Armenia and colleagues chose to highlight newly statistically significant genes in the long tail for which—despite their low prevalence—there was known molecular biology supporting a strong and clear role in cancer. The frequent elevation of these mutations to detectably high cancer-cell counts in tumors (on the infrequent occasions when the mutations occur in a suitable cancer stem-cell lineage) indicated by intensity of selection validates the inference Armenia and colleagues made that they are genes with clear and important roles in the cancers that carry those mutations. Reciprocally, we would argue that it also points out how useful it is to quantify the intensity of selection within cancer lineages, and how useful it is to "wag" the long tail of oncogenic drivers—ranking the genes by their cancer effect size rather than by their *P* value or their prevalence.

Supporting information

S1 Table. Mutation rates, selection intensities, and prevalences for sustitutions.
(TXT)

References

1. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, et al. The long tail of oncogenic drivers in prostate cancer. Nat Genet. 2018; https://doi.org/10.1038/s41588-018-0078-z PMID: 29610475
2. Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, et al. Genomic hallmarks of localized, non-indolent prostate cancer. Nature. 2017; 541: 359–364. https://doi.org/10.1038/nature20788 PMID: 28068672
3. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505: 495–501. https://doi.org/10.1038/nature12912 PMID: 24390350
4. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007; 82: 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x PMID: 17944619
5. Cannataro VL, Gaffney SG, Stender C, Zhao Z, Philips M, Greenstein AE, 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: 2444–2455. https://doi.org/10.1038/s41388-017-0105-z PMID: 29453361
6. Cannataro VL, Gaffney SG, Townsend JP. Effect sizes of somatic mutations in cancer. Journal of the National Cancer Institute. 2018; 110: 1171–1177. https://doi.org/10.1093/jnci/djy168 PMID: 30365005
7. Cannataro VL, Townsend JP. Neutral theory and the somatic evolution of cancer. Mol Biol Evol. 2018; https://doi.org/10.1093/molbev/msy079 PMID: 29684198
8. Ryk C, de Verdier P, Montgomery E, Peter Wiklund N, Wiklund F, Grönberg H. Polymorphisms In The Nitric-Oxide Synthase 2 Gene And Prostate Cancer Pathogenesis. Redox Biol. 2015; 5: 419.

9. Wang T, Chen Y-H, Hong H, Zeng Y, Zhang J, Lu J-P, et al. Increased nucleotide polymorphic changes in the 5'-untranslated region of delta-catenin (CTNND2) gene in prostate cancer. Oncogene. 2009; 28: 555–564. https://doi.org/10.1038/onc.2008.399 PMID: 18978817

10. Yang X, Zhang Y, Hosaka K, Andersson P, Wang J, Tholander F, et al. VEGF-B promotes cancer metastasis through a VEGF-A–independent mechanism and serves as a marker of poor prognosis for cancer patients. Proc Natl Acad Sci U S A. National Academy of Sciences; 2015; 112: E2900–E2909. https://doi.org/10.1073/pnas.1503500112 PMID: 25991856