TECHNICAL RESPONSE

MUTATION DETECTION

Response to Comment on "DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification"

Lixin Chen, Pingfang Liu, Thomas C. Evans Jr.*, Laurence M. Ettwiller*

Following the Comment of Stewart *et al*., we repeated our analysis on sequencing runs from The Cancer Genome Atlas (TCGA) using their suggested parameters. We found signs of oxidative damage in all sequence contexts and irrespective of the sequencing date, reaffirming that DNA damage affects mutation-calling pipelines in their ability to accurately identify somatic variations.

Previously, we devised a metric termed the global imbalance value (GIV) to evaluate how mutagenic damage affects sequencing accuracy (*1*). We showed that mutagenic damage is pervasive in public sequencing datasets and confounds the identification of somatic variants with low to moderate (1 to 5%) allelic frequency. Following our publication, the principle of global imbalance was incorporated by the International Cancer Genome Consortium (ICGC) as one of five measures used to construct a quality rating for each cancer genome (*2*). Stewart *et al*. (*3*) reaffirm the presence of sequencing errors derived from damage in raw sequencing reads (*1*, *4*–*6*), but they question the relevance of these errors in downstream analysis. Many of their points can be distilled down to whether (i) sequencing datasets with a GIV score greater than 2 can affect the output of mutation callers, and (ii) measures have been taken since 2012 to mitigate these errors. Stewart *et al*. also claim (iii) that the default cutoff chosen in our paper led to incorrect conclusions related to the context specificity of oxidative damage–induced errors. We address these points in order.

i) DNA damage affects mutation-calling pipelines: Responding to whether damage affects downstream analysis, we examined whether damage-induced errors can be detected in mutation calls generated using standard approaches. For this, we downloaded VCF files generated in 2016 using MuTect (*7*) from the TCGA data portal (https://portal.gdc.cancer.gov/) and examined whether the fraction of G:C > T:A point mutations can be correlated with the estimated level of damage in the corresponding sequencing runs.

We observed an overall increase in the fraction of G:C > T:A mutations from a median of

New England Biolabs Inc., Ipswich, MA 01938, USA.
***Corresponding author. Email: evanst@neb.com (T.C.E.); ettwiller@neb.com (L.M.E.)**

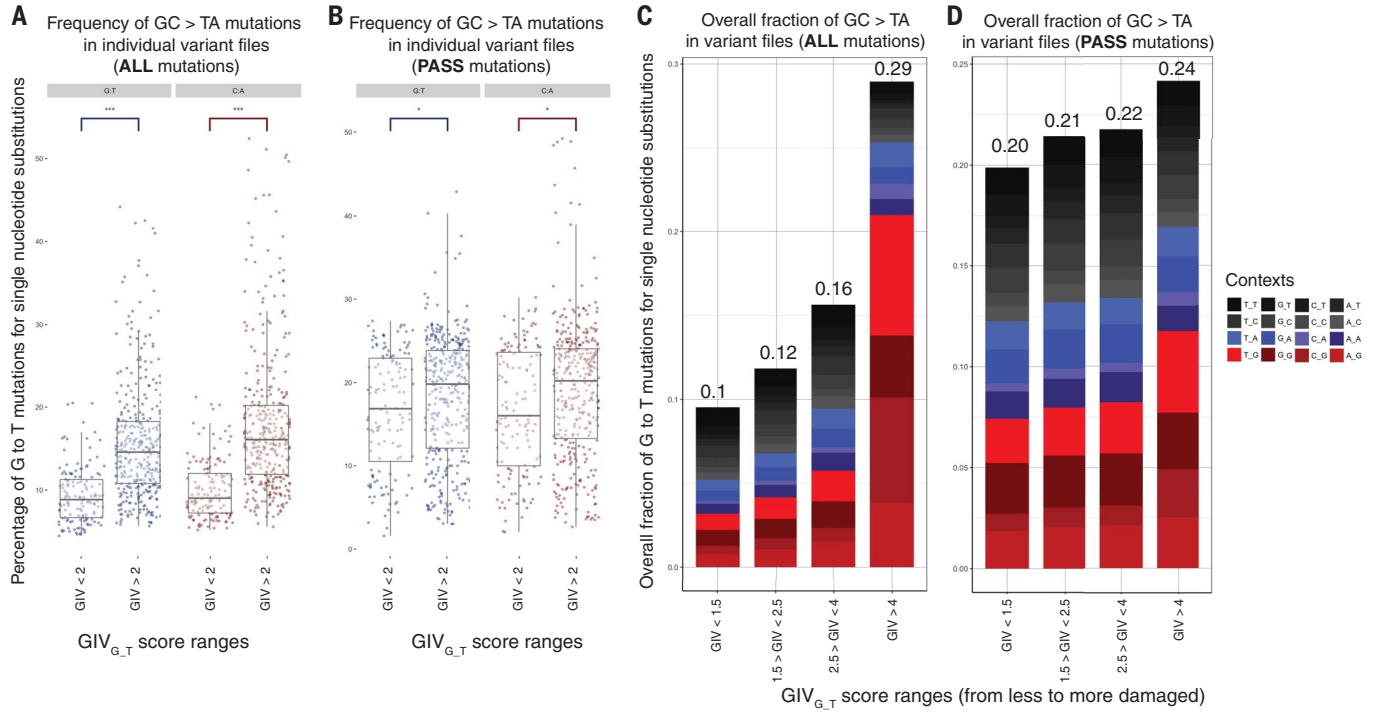

Fig. 1. TCGA mutation calls are affected by DNA damage. (**A**) Percentage of G > T and C > A point mutations relative to the total number of point mutations in variant files downloaded from the TCGA data portal. Two groups are shown: group 1, for which the sequencing of the cancer sample leads to a GIV score of <2, and group 2, for which the sequencing data have a GIV score of >2 (damaged samples). Damaged samples show a significantly elevated fraction of G-to-T mutations ($P < 2.2 \times 10^{-16}$, Wilcox test). (**B**) Same as (A) using only the mutations labeled as PASS. Even for PASS mutations, samples that are damaged show an elevated fraction of G-to-T mutations [$P = 0.03$ (G_T), $P = 0.012$ (C_A), Wilcox test]. *$P \le 0.05$, ***$P \le 0.001$. (**C**) Overall fraction of G-to-T point mutations relative to the total number of point mutations in the ALL-MuTect variant files downloaded from TCGA for cancer samples with $GIV_{G\_T} < 1.5$, $1.5 < GIV_{G\_T} < 2.5$, $2.5 < GIV_{G\_T} < 4$, and $GIV_{G\_T} > 4$. Mutations are color-coded according to sequence context. (**D**) Same as (C) except using only the mutations labeled as PASS. GIV scores were calculated using the cutoff (Q score > 20) as in Stewart *et al*.

8.9% in cancer samples with no or low levels of damage ($GIV_{G\_T} < 2$) to a median of 15.3% in samples that were damaged ($GIV_{G\_T} > 2$) (Fig. 1A). Restricting the analysis to only the MuTect PASS mutations (*7*) that represent the most confident set of mutations, we found an overall increase in the fraction of G:C > T:A mutations from a median of 16.8% to a median of 20.0% in damaged samples (Fig. 1B). This increase in the fraction of G:C > T:A mutations for damaged samples is observable even for moderately damaged samples ($1.5 < GIV_{G\_T} < 2.5$) (Fig. 1, C and D). Collectively, this demonstrates that standard mutation callers are affected by damage-induced errors, even in analysis pipelines from 2016.

ii) Damage can be found post-2012: Stewart *et al.* assert that the effects of oxidative damage on sequencing data were known and mitigated after 2012 (*3*). Although we cite the study accordingly, we note that they did not consider broad mutagenic damage in key public databases, nor is their claim of full mitigation supported by the mean GIV scores from samples of TCGA datasets from 2010 to 2015 (Table 1). Our findings indicate that artifacts consistent with DNA damage are present in these databases irrespective of the sequencing date.

iii) G-to-T transversion consistent with 8-oxo-deoxyguanosine (8-oxo-dG) damage can be found in all sequence contexts: Stewart *et al.* found that the CCG > CAG context specificity is a feature of oxidative damage of dG and is essential for mutation signature analysis (*4*). In contrast, we reported that a G-to-T imbalance indicative of damage is present regardless of the preceding or succeeding nucleotide. Stewart *et al.* claim that this divergence was due to a default cutoff (base quality threshold Q > 30) (*8*, *9*), which would eliminate most of the data in highly damaged samples, and they conclude that the oxidative damage of dG lacks context specificity. We therefore used the cutoff suggested by Stewart *et al.*, reanalyzed our previously reported in-house sequencing runs, and again found evidence of damage-induced errors in all sequence contexts (Fig. 2A).

Extending our study to public sequencing datasets using a more quantitative measurement of damage, we observed an increase in the overall fraction of G-to-T transversions in damaged samples irrespective of sequence context (Fig. 2B). Furthermore, some samples had a greater increase in the fraction of G-to-T transversions when a purine was located 3′ to the dG. Thus, the CCG > CAG context observed by Stewart *et al.* represents only a subset of errors from oxidative damage.

We value the cross-examination of our scientific work, as it has strengthened some of our earlier statements and refined others. Damage-induced sequencing errors are pervasive; we estimated that for 73% of TCGA sequencing runs analyzed in our paper, >50% of G-to-T raw read variants correspond to damage ($GIV_{G\_T} > 2$). Stewart *et al.* assert that we have misled readers to believe that

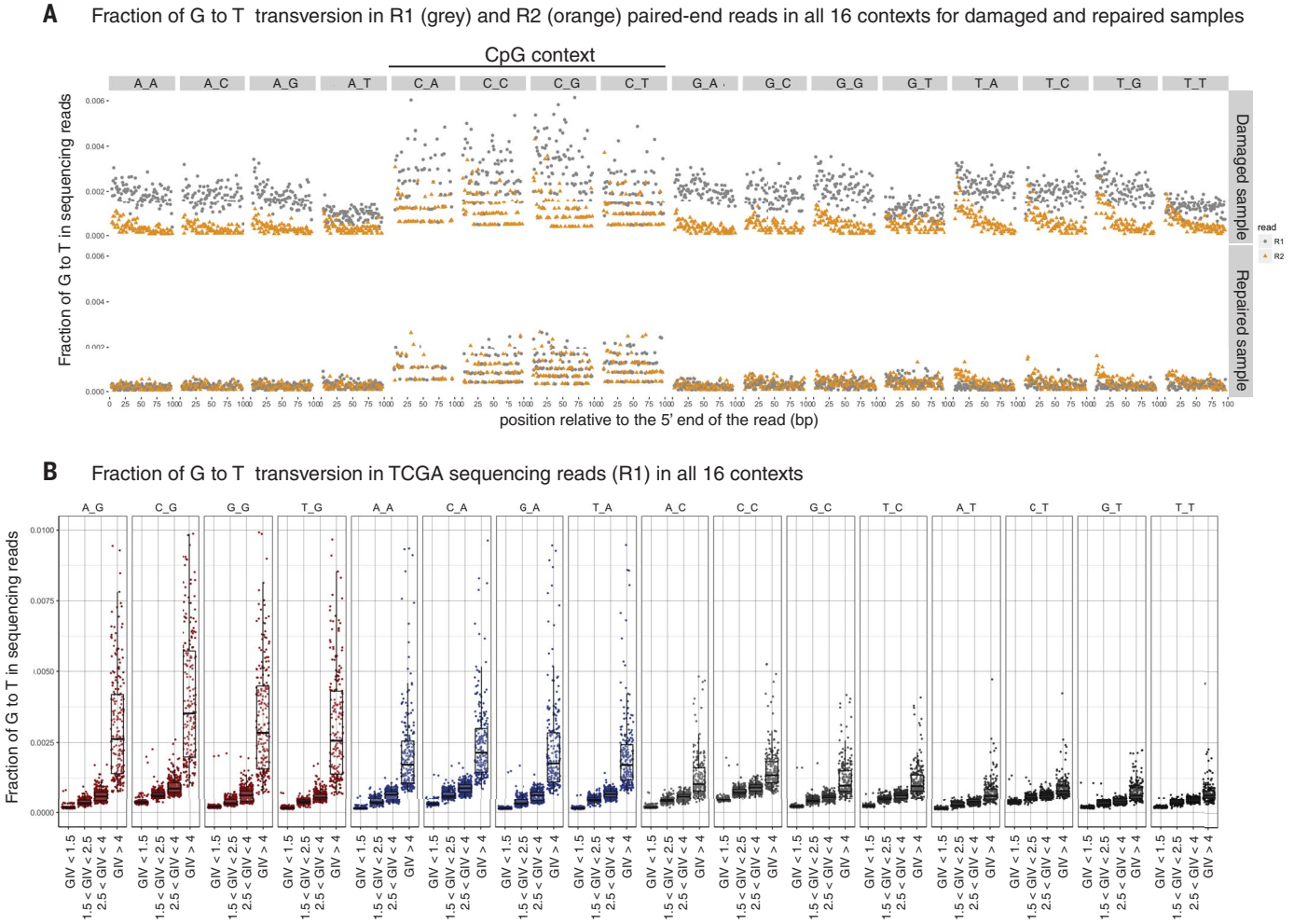

Fig. 2. G > T transversions in all 16 three-base sequence contexts. (**A**) Reanalysis of the sequencing dataset using the GIV scores calculated using the suggested cutoff (Q score > 20) by Stewart *et al*. Read variant frequencies of G to T in R1 (gray) and R2 (orange) as a function of the position were plotted relative to the 5′ end of the read. The upper panel shows the damaged sample [(*1*), figure S1, Water]; the lower panel corresponds to the same sample repaired with damage repair enzymes [(*1*), figure S1, Water + repair]. A line denotes a different error pattern at CpG contexts due to the limited amount of data at these sites. (**B**) Quantitative measurement of the fraction of G to T transversions in sequencing reads from a subset of TCGA datasets in all 16 three-base sequence contexts (N_N denotes NGN > NTN context). Red and blue denote 3′ purine context (NGG > NTG and NGA > NTA contexts, respectively). The datasets were grouped into four categories, from no damage ($GIV_{G\_T} < 1.5$) to severely damaged samples ($GIV_{G\_T} > 4$).

Table 1. GIV_{G_T} scores for sequencing runs performed between 2010 and 2015. We reanalyzed our data using the date printed on the BAM file header and calculated the mean, median, and first and third quantiles of GIV scores for each sequencing year. Although we identify an improvement in 2013, sequencing performed in 2014 and 2015 had an average GIV score comparable to the average GIV score obtained in and prior to 2012. GIV scores were calculated with the suggested cutoff (Q score > 20).

Year	First quantile	Median	Mean	Third quantile
2010	2.307	2.414	2.371	2.545
2011	2.606	3.026	4.586	5.199
2012	2.026	2.527	4.069	3.748
2013	1.219	1.288	1.678	2.136
2014	4.14	4.361	4.313	4.569
2015	2.516	2.876	2.749	3.09

this level of damage is extensive, based on the fact that damage accounts for only a fraction of the overall error rate. However, contrary to intrinsic errors, damage-induced errors such as 8-oxo-dG consistently produce the same error type (in this case, G to T) even at high Q scores. This distinction between the intrinsic error of the sequencing instrument and the error rate at high Q scores is particularly important because mutation callers rely on Q scores to filter out false-positive mutations. Thus, treating the impact of damage as if it were part of the overall sequencing error rate underestimates the effect of damage.

As part of our analysis, we required a metric to correctly estimate the percentage of false-positive variants relative to the total number of variants. This metric, which we termed the rate of false positives, is described in the supplementary materials of Chen *et al.* and does not require knowing the false and true negatives. Its name, however, seems to have caused confusion, and in hindsight a different name could have been chosen.

Finally, the purpose of the GIV score is to provide an estimate of the fraction of raw read variants that are derived from damage. As exemplified by the ICGC, such measures can be used to flag samples of poor quality for improved downstream analysis. Our study does not call into question previous studies that used data from TCGA, as stated by Stewart *et al.* Instead, we agree with Stewart *et al.* that DNA damage can affect sequencing accuracy and downstream analysis. Undoubtedly, TCGA and similar public databases are valuable resources, and we hope that this dialogue increases the awareness of possible sources of error and mitigation strategies.

REFERENCES AND NOTES

1. L. Chen, P. Liu, T. C. Evans Jr., L. M. Ettwiller, *Science* **355**, 752–756 (2017).
2. J. P. Whalley *et al.*, BioRxiv 140921 [Preprint]. 19 September 2017.
3. C. Stewart, I. Leshchiner, J. Hess, G. Getz, *Science* **361**, eaas9824 (2018).
4. M. Costello *et al.*, *Nucleic Acids Res.* **41**, e67 (2013).
5. M. W. Schmitt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14508–14513 (2012).
6. A. M. Newman *et al.*, *Nat. Biotechnol.* **34**, 547–555 (2016).
7. K. Cibulskis *et al.*, *Nat. Biotechnol.* **31**, 213–219 (2013).
8. Illumina, "Quality Scores for Next-Generation Sequencing"; www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf.
9. Current Illumina sequencing platforms have a reported minimum of 87% of bases with a Q score above 30, indicating that a cutoff of Q = 30 is not aggressive and does not significantly bias the dataset.

ACKNOWLEDGMENTS

We thank H. Runz for useful comments. **Funding:** Supported by New England Biolabs Inc. **Author contributions:** L.M.E. performed the data analysis; L.C., P.L., and T.C.E. contributed ideas and participated in writing the manuscript. **Competing interests:** The authors are listed as inventors on U.S. provisional serial number 62/376,165, submitted by New England Biolabs Inc., which covers improved sequence accuracy determination of a nucleic acid sample. **Data availability:** All data are referenced or available in the manuscript or the supplementary materials.

6 April 2018; accepted 29 August 2018
10.1126/science.aat0958

Response to Comment on "DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification"

Lixin Chen, Pingfang Liu, Thomas C. Evans Jr. and Laurence M. Ettwiller

Science **361** (6409), eaat0958.
DOI: 10.1126/science.aat0958