TECHNICAL COMMENT

MUTATION DETECTION

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

Chip Stewart[1]*, Ignaty Leshchiner[1]*, Julian Hess[1], Gad Getz[1,2,3]†

Chen *et al*. (Reports, 17 February 2017, p. 752) highlight an important problem of sequencing artifacts caused by DNA damage at the time of sample processing. However, their manuscript contains several errors that led the authors to incorrect conclusions. Moreover, the same sequencing artifacts were previously described and mitigated in The Cancer Genome Atlas and other published sequencing projects.

The sequencing artifacts discussed in Chen *et al*. (*1*) have been described in publications from The Cancer Genome Atlas (TCGA) and other cancer genome projects (*2*–*4*). Accordingly, effective mitigation strategies have long been implemented, including software pipelines and improved library preparation methods, as described in Costello *et al*. (*2*). Thus, findings described in TCGA publications (*5*–*9*) were not affected by oxidative damage, as suggested by Chen *et al*. The authors do raise general awareness of sequencing artifacts, which include machine errors (*2*), DNA oxidation (*2*–*4*), and DNA cross-linking (in clinically used formalin-fixed tissues) (*10*) among others; however, their paper contains several errors that led to incorrect conclusions. The errors affect the following:

i) Estimation of oxidative damage levels: Although their reported oxidative damage (8-oxo-G) metric, GIV_{G_T}, is essentially equivalent to an earlier reported metric, oxoQ (*2*–*4*, *11*), their software implementation is limited by using reads aligned only to the forward strand and biased by filtering out low-quality bases. OxoQ was designed to be interpreted as a (Phred-like) base quality score. As such, oxoQ can be compared to typical levels of sequencing errors, which for most Illumina sequencing protocols is roughly at the level of Q ~ 30 (i.e., an error rate of 1 per 1000 bases). We initially noticed a poor agreement between the oxoQ and GIV_{G_T} metrics for damaged samples (oxoQ < 30; Fig. 1A), whereas scores were consistent for samples with low levels of DNA damage. Examining the methods of Chen *et al*., we found that the authors' code with default parameters applied a high base quality threshold (Q > 30) and removed sites where total coverage exceeds

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02124, USA. [2]Massachusetts General Hospital Cancer Center and Department of Pathology, Boston, MA 02114, USA. [3]Harvard Medical School, Boston, MA 02115, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: gadgetz@broadinstitute.org

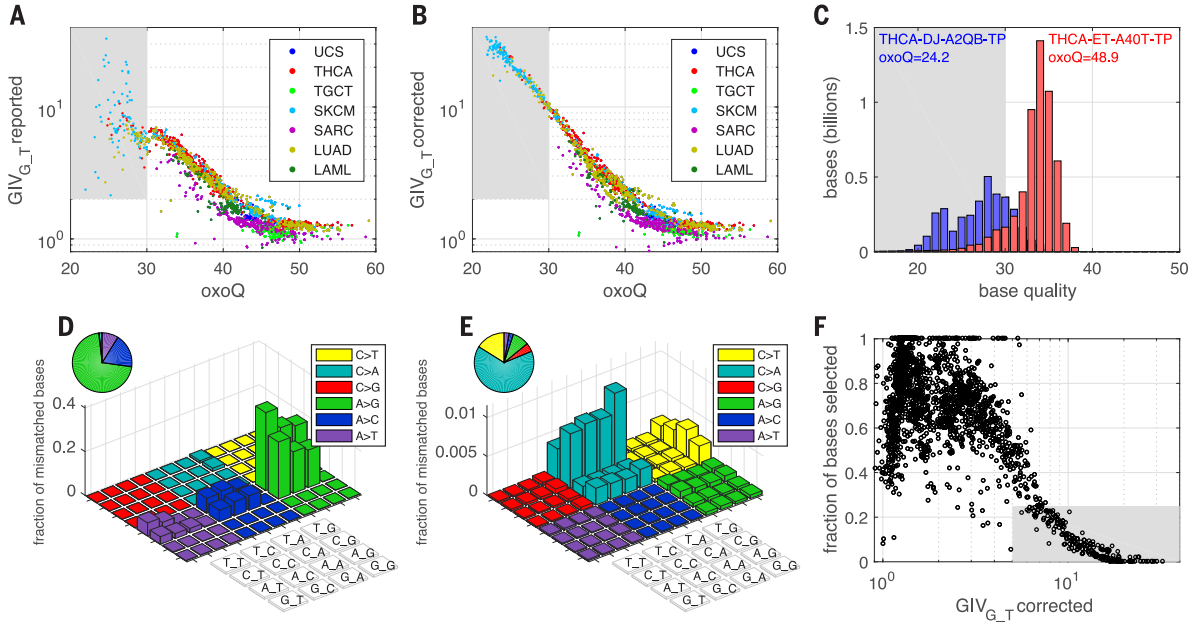

Fig. 1. Comparison of GIV_{G_T} to oxoQ score. (**A**) Picard oxoQ score versus GIV_{G_T} for ~1900 TCGA tumor exomes. Lack of agreement between oxoQ and GIV_{G_T} scores is apparent in samples with high DNA damage (oxoQ < 30) highlighted in the gray box. GIV_{G_T} is calculated with Chen *et al*.'s code (estimate_damage.pl) with default parameters; oxoQ was calculated with Picard CollectSequencingArtifactMetrics (*11*) output. (**B**) OxoQ versus GIV_{G_T} corrected (using base quality Q > 20 and coverage depth ≥20×, comparable to Picard defaults) showing excellent agreement. Color code in (A) and (B) indicates data from seven different TCGA tumor types. (**C**) Histogram of individual base qualities in a typical TCGA 8-oxo-G–damaged sample (oxoQ = 24.2) compared to a low-damage sample (oxoQ = 48.9). Most of the bases in the damaged sample have Q < 30 and hence are ignored by default GIV_{G_T}. This explains both the lack of agreement between the scores seen in (A) and the authors missing the sequence context associated with 8-oxo-G damage [(D) versus (E)]; that is, the Q > 30 filtering removed bases with the type of damage that is being quantified. (**D** and **E**) "Lego" plot showing the distribution of errors in different 5′ and 3′ sequence contexts (plotted in reverse complement by convention), with Chen *et al*.'s (D) and our corrected (E) filtering criteria, showing distortion from the characteristic 8-oxo-G context pattern arising from data selection criteria. (**F**) The fraction of bases that remain after applying Chen *et al*.'s filters, showing that most bases are removed in highly damaged samples.

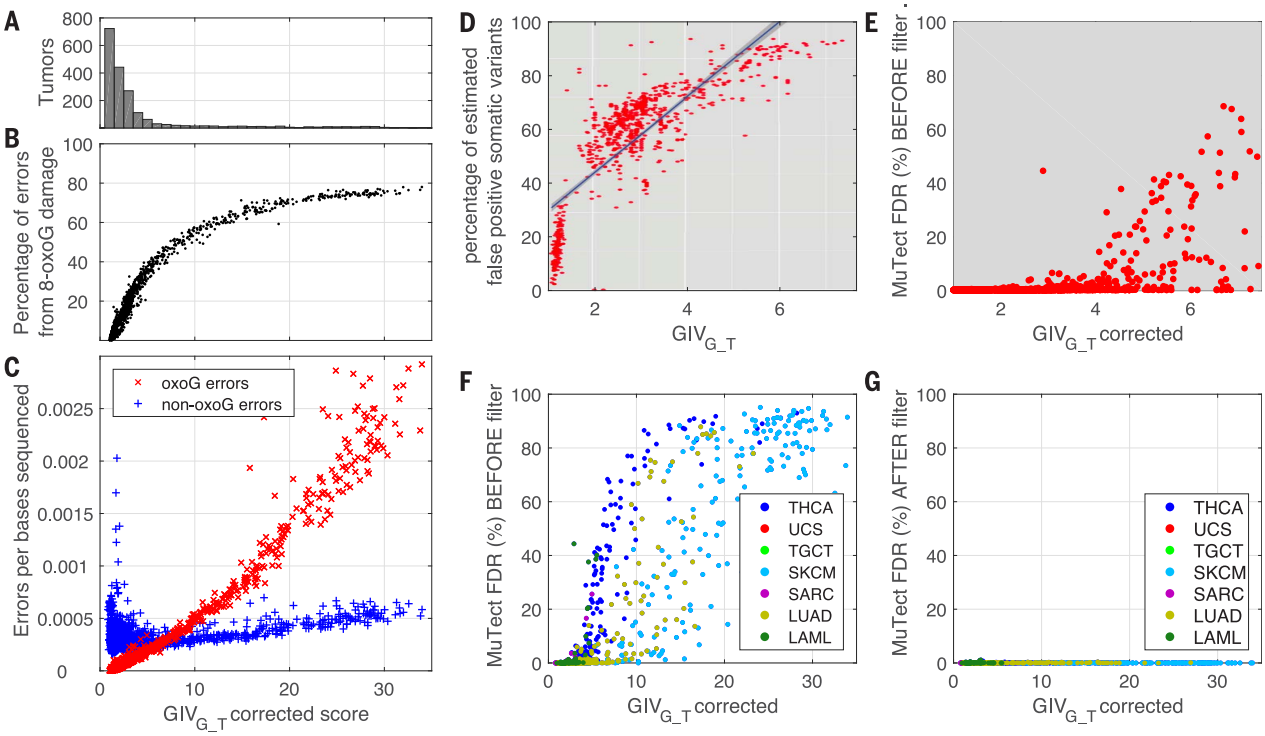

Fig. 2. Comparison of 8-oxo-G–related error rates and other error modes. (**A** to **C**) Samples with corrected GIV_{G_T} damage levels below 5 (equivalent to oxoQ ~ 35) have fewer 8-oxo-G–related base errors (red) relative to other sequencing error modes (blue). (A) Count of tumors with a specific corrected GIV_{G_T} score (the majority of TCGA samples have very low damage levels). (B) Percentage of additional base mismatches caused by 8-oxo-G in relation to GIV_{G_T} score. (C) 8-oxo-G–related error rates (red) and other error modes (blue). Other modes dominate until corrected GIV_{G_T} ~ 5. (**D**) Copy of figure 4E from Chen *et al*. (*1*). (**E**) Our estimated FDR versus corrected GIV_{G_T} for ~1900 TCGA tumors from MuTect (*14*) before 8-oxo-G filtering (GIV_{G_T} < 7.5). (**F** and **G**) Estimated FDR versus corrected GIV_{G_T} for ~1900 TCGA tumors from MuTect before (F) and after (G) 8-oxo-G filtering [as described in (*2*)]. Results demonstrate no inflation in number of mutation calls after applying the 8-oxo-G software filter.

100×, which discarded the bulk of the data for the most damaged samples (Fig. 1, C to F). Adjusting the filtering criteria to include bases with Q > 20 and to allow sites with depth above 100× restored concordance between GIV_{G_T} and oxoQ scores (Fig. 1B).

ii) Claim of no context specificity of 8-oxo-G damage: The authors' conclusion that they "did not observe nucleotide context specificity" of 8-oxoG damage (*1*) is a consequence of their filtering out most of the supporting data, as described above. Sequence context specificity is a key feature of 8-oxo-G damage (*12*), can affect mutational signature analysis, and can be visualized using "Lego" plots (Fig. 1, D and E). These plots show the error rate of each type of base substitution in its three-base sequence context (*2*). The damaged bases reported by Chen *et al*.'s script (Fig. 1D) have a severely distorted error profile inconsistent with the previously reported (*12*) 8-oxo-G damage pattern (Fig. 1E, C > A errors in the Lego plot with a clear peak at the sequence context CCG > CAG), whereas the corrected GIV_{G_T} script [and the previous implementation in (*2*)] results in a distribution of errors (Fig. 1E) consistent with 8-oxo-G damage. Thus, 8-oxo-G damage does in fact have a sequence context, which the authors apparently missed as a result of their filtering strategy.

iii) Interpretation of 8-oxo-G damage levels: The authors state that "73% of the TCGA sequencing runs showed extensive damage, with a GIV_{G_T} > 2." A GIV_{G_T} metric of 2 indicates that the G > T error rate in the 8-oxo-G mode is twice the error rate of the non–8-oxo-G mode (background rate for G > T errors). Even if the 8-oxo-G artifacts are twice the context-specific background level (i.e., GIV_{G_T} = 2), this corresponds to only a 5 to 10% increase in the overall base-level error rate (summed over all sequence contexts; Fig. 2, A to C), which is less than the intersample variability of error rates at a fixed oxoQ. A 5% increase in the base-level error rate results in a minor, if any, increase in false-positive mutation calls (Fig. 2, E and F), because calling algorithms are designed to handle typical levels of sequencing error. Only at $GIV_{G\_T} \gtrsim 5$ (equivalent to oxoQ $\lesssim$ 35) do the additional errors from 8-oxo-G become comparable to the sum of all other errors and have an adverse impact on variant calling. The vast majority of samples in TCGA exhibit only minor 8-oxo-G damage that has minimal impact on mutation calling. Consequently, the claim that 73% of TCGA sequencing runs have extensive damage is misleading.

iv) Estimation of the false positive rate (FPR) of mutation calling: The authors define an FPR metric, which suggests that mutation calls for many samples in public databases contain >50% falsely detected somatic variants due to sequencing artifacts (Fig. 2D). However, their metric does not reflect the actual somatic mutations used for analyses in (*3*–*10*) and other TCGA publications [e.g., a recent reanalysis of TCGA data (*13*)], but instead represents candidate variants supported by as few as two nonreference reads [not considered as somatic variants by most mutation callers (*14*)]. Moreover, the mathematical definition of their FPR metric is neither a FPR nor a false discovery rate (FDR), which is highlighted by the fact that it can range between –1 and 1 (and not between 0 and 1). Therefore, Chen *et al*. incorrectly concluded that the FPR exceeds 0.5 in samples that have a low level of DNA damage with nearly no damage-induced false-positive mutation calls (Fig. 2, D to G).

v) Finally, Chen *et al*. overlook publications describing 8-oxo-G damage (*2*–*10*) that TCGA and other projects have mitigated with laboratory protocols and software filtering strategies when sequencing library regeneration was impractical (Fig. 2, F and G). Moreover, they claim that "recent submissions to TCGA (November to December 2015) displayed similar G-to-T imbalances" but mistakenly interpreted the time that the data repository was last updated as the time of sequencing data generation (repository updates occurred periodically for unrelated issues). The actual dates of generating the sequencing data should be obtained from the sequencing metadata. In fact, most TCGA samples sequenced after

October 2012 had low 8-oxo-G damage due to improved library preparation [as described in (*2*)].

In conclusion, we disagree with the assessment of Chen *et al.* regarding the quality of published cancer genome projects. Raw sequencing data inherently contain errors; therefore, to avoid misinterpreting the data, it is important that researchers use established procedures and carefully curated datasets in downstream analyses.

REFERENCES AND NOTES

1. L. Chen, P. Liu, T. C. Evans Jr., L. M. Ettwiller, *Science* **355**, 752–756 (2017).
2. M. Costello *et al.*, *Nucleic Acids Res.* **41**, e67 (2013).
3. T. J. Pugh *et al.*, *Nat. Genet.* **45**, 279–284 (2013).
4. B. D. Crompton *et al.*, *Cancer Discov.* **4**, 1326–1341 (2014).
5. Cancer Genome Atlas Research Network, *Cell* **159**, 676–690 (2014).
6. Cancer Genome Atlas Research Network, *N. Engl. J. Med.* **372**, 2481–2498 (2015).
7. Cancer Genome Atlas Research Network, *Cell* **161**, 1681–1696 (2015).
8. Cancer Genome Atlas Research Network, *Nature* **541**, 169–175 (2017).
9. Cancer Genome Atlas Research Network, *Cancer Cell* **32**, 185–203.e13 (2017).
10. M. Giannakis *et al.*, *Cell Rep.* **17**, 1206 (2016).
11. Picard; https://broadinstitute.github.io/picard.
12. Y. Margolin, V. Shafirovich, N. E. Geacintov, M. S. DeMott, P. C. Dedon, *J. Biol. Chem.* **283**, 35569–35578 (2008).
13. K. Ellrott *et al.*, *Cell Syst.* **6**, 271–281.e7 (2018).
14. K. Cibulskis *et al.*, *Nat. Biotechnol.* **31**, 213–219 (2013).

ACKNOWLEDGMENTS

We are grateful for useful comments from E. S. Lander, M. Costello, N. J. Lennon, and L. Lichtenstein as well as editing help from M. Miller. **Funding:** Partially supported by NIH grant 1U24CA210999. **Author contributions:** C.S., I.L., J.H., and G.G. conceived the analysis, carried out the analysis, and wrote the text. **Competing interests:** The authors have no competing interests. **Data availability:** Raw data are available via TCGA controlled access, https://tcga-data.nci.nih.gov/docs/publications/tcga/accesstiers.html; code and tables are available at https://github.com/broadinstitute/damage.

11 January 2018; accepted 29 August 2018
10.1126/science.aas9824