REPORT

MUTATION DETECTION

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

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

Mutations in somatic cells generate a heterogeneous genomic population and may result in serious medical conditions. Although cancer is typically associated with somatic variations, advances in DNA sequencing indicate that cell-specific variants affect a number of phenotypes and pathologies. Here, we show that mutagenic damage accounts for the majority of the erroneous identification of variants with low to moderate (1 to 5%) frequency. More important, we found signatures of damage in most sequencing data sets in widely used resources, including the 1000 Genomes Project and The Cancer Genome Atlas, establishing damage as a pervasive cause of sequencing errors. The extent of this damage directly confounds the determination of somatic variants in these data sets.

Genomic variations in somatic cells can result in disease states, including cancer (*1*–*3*). Thus, accurate tumor-associated variant detection, which may help direct personalized treatments, is important for cancer diagnosis and prognosis. Next generation sequencing (NGS) has revolutionized variant identification and characterization. Nonetheless, owing to tumor heterogeneity and/or contamination by normal cells, somatic cancer variants are often found at low allelic frequencies (*4*, *5*), confounding their identification.

Detection of low allelic frequency variants is achieved through deep sequencing and specialized data analysis algorithms that detect variants in a limited number of reads. Data analysis is challenged by artifactual errors that display the same low allelic frequency as cancer mutations, with the level of artifactual errors defining the threshold for low allelic variant detection. Most sequencing errors are thought to result from polymerase chain reaction mistakes or sequencing miscalls (*6*). Meanwhile, mutagenic DNA damage is recognized as a major source of sequencing errors only in specialized samples—for example, formalin-fixed paraffin-embedded (*7*), ancient (*8*), and circulating tumor DNA (*9*). Furthermore, another study demonstrated that library preparation induces oxidative damage (*10*), raising the possibility that sequencing high-quality human genomic DNA may also be affected by mutagenic damage.

We explored this possibility by measuring damage in sequencing runs. For this, we used the fact that mutagenic damage leads to a global imbalance between variants detected in read 1 (R1) and read 2 (R2) in paired-end sequencing (Fig. 1A) (*11*). The degree of this imbalance directly correlates with the amount of damage present in a sample. We devised an analysis strategy based on this imbalance to deconvolute both the origin and orientation of variants and computed a metric, the Global Imbalance Value (GIV) score, that is indicative of damage (*11*).

The algorithm produces 12 GIV scores, one per variant type. Here, a GIV score above 1.5 is defined as damaged. At this GIV score, there are 1.5 times more variants on R1 than on R2, suggesting that at least one-third of the variants are erroneous. Undamaged DNA samples have a GIV score of 1. To experimentally validate the GIV score and provide an independent damage quantification, we used human genomic DNA containing various amounts of 7,8-dihydro-8-oxoguanine (8-oxo-dG), resulting in G-to-T transversions after amplification (*10*, *12*). We also treated the damaged DNA with an enzyme cocktail that repairs DNA damage before library preparation (*11*, *13*). Sequencing the same sample with and without DNA repair enzyme treatment quantified the rate of erroneous variants specifically introduced by damage. Confirming previous findings (*10*), the G-to-T transversion frequency varied according to library preparation conditions (figs. S1 and S2) (*11*). Notably, excess G-to-T variants were only observed in R1 sequences, whereas C-to-A variants were in excess in R2 sequences, leading to a GIV_{G_T} score > 1 (Fig. 1B). Repair enzyme treatment abolished this imbalance and reduced the GIV_{G_T} score to 1. The GIV score correlated with the variant excess measured experimentally (fig. S1E), demonstrating that the GIV score can be used to accurately estimate the extent of damage in publicly available data sets. We estimated that the GIV score calculation is accurate at >2 million reads (fig. S3B).

To estimate the extent of damage in public data sets, we determined the GIV scores of individual sequencing runs from the 1000 Genomes Project (*14*) and a subset of The Cancer Genome Atlas (TCGA) data set (*11*). Both data sets showed widespread damage, particularly those leading to an excess of G-to-T variants (Fig. 2). Specifically, 41% of the 1000 Genomes Project data sets had a GIV_{G_T} score ≥ 1.5, indicative of damaged samples (Fig. 2A). Furthermore, 73% of the TCGA sequencing runs showed extensive damage, with a GIV_{G_T} > 2. This indicates that the majority of G-to-T observations are erroneous and establishes damage as a pervasive cause of errors in these data sets (Fig. 2B). Further, we found no nucleotide context specificity of G-to-T imbalances in these data sets (fig. S4). Additionally, an A to T imbalance (fig. S5) leading to GIV_{T_A} > 1.5 and a C to T imbalance (GIV_{C_T} > 1.5) were detected in 0.5 and 3% of the TCGA data set, respectively (*11*). Finally, recent submissions to TCGA (November to December 2015) displayed similar G-to-T imbalances and accentuated A-to-T imbalances (fig. S6). These results confirm that most publicly available data sets, including recent submissions to TCGA, have signatures of damage leading to erroneous calls in at least one-third of the G-to-T variant reads.

Our data showed damage leading to G-to-T transversions to be stochastic (fig. S1A) (*11*). Such stochasticity implies that errors derived from damage are expected to be present at low allelic fractions. Therefore, the identification of low-frequency variants—e.g., somatic variants—would be affected by damage, whereas variants present at higher frequency—e.g., germline variants—would be unaffected. To evaluate how damage affects somatic variant identification, we repeated the oxidative damage experiments using common library preparation procedures. We further performed target enrichment using a commercial cancer panel probe set to achieve high sequencing depth of 151 annotated cancer genes (Fig. 3A) (*11*).

Candidate variants were grouped according to frequency, with very low (<1%), low to moderate (1 to 5%), medium (6 to 10%) and high (>10%) frequency classes. We found that DNA repair eliminates 77 and 82% of G-to-T and C-to-A variant positions in the very low and low to moderate frequency classes, respectively, indicating that those positions were erroneous and due to damage (Fig. 3B) (*11*). Notably, most candidate variant positions in the low- to moderate-frequency class were due to damage despite harboring multiple evidences of variant reads (≥3). The imbalance of G-to-T compared with C-to-A positions in the unrepaired data set (Fig. 3C) confirms the role of damage in erroneous variant calling. In the 0.79-Mb region included in the cancer panel, we found 195 genomic locations with low to moderate G-to-T and C-to-A variants, with 50 marked as deleterious and 7 annotated as nonsense, according to PredictSNP2 (*15*) (table S2). In comparison, the repaired data set contained only 12 genomic locations with low to moderate G-to-T and C-to-A variants.

These results indicate that more than 180 positions are false positives and are directly confounding the identification of real somatic variants.

New England Biolabs Inc., 240 County Road, Ipswich, MA 01938-2723, USA.
*Corresponding author. Email: evanst@neb.com (T.C.E.), ettwiller@neb.com (L.M.E.)

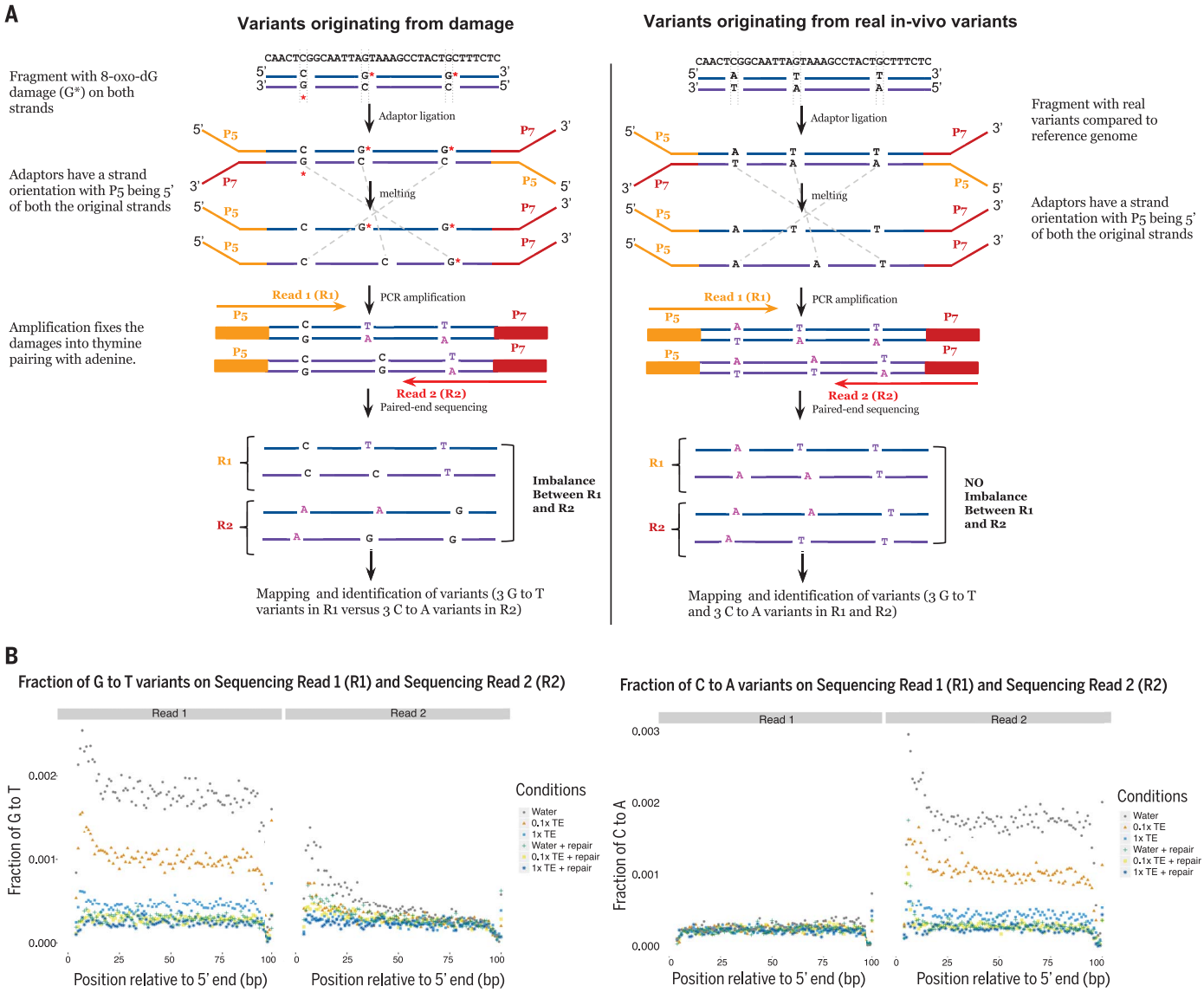


Fig. 1. GIV score. (**A**) GIV score principle. Illumina sequencing adaptors P5 and P7 are directional in nature, enabling consistent paired-end sequencing within clusters. This property results in sequencing of the original strand orientation in the R1 reads (from the P5 adaptor), whereas the reverse complement orientation is read in the R2 reads (from the P7 adaptor). Because damage affects only one base of a pair, damage such as 8-oxo-dG leads to an excess of G-to-T transversion errors when R1 is mapped to a reference genome, whereas, the R2 reads will show an excess of the reverse complement of G-to-T—i.e., C-to-A errors—instead. As a consequence, there is a global imbalance in the number of G-to-T variants in R1 compared with R2 sequences. This imbalance is specific to damage and is the basis of the GIV score (left panel) (*11*). Contrasting with damage, true variations lead to no imbalance (right panel). (**B**) Variant profile. The fraction of G-to-T (left panel) and C-to-A variants (right panel) in R1 and R2 sequences were plotted as a function of the read (R1 or R2) and position, in base pairs (bp). Acoustic shearing in different solutions generated various levels of G-to-T in R1 or C-to-A in R2. In all cases, treatment of the DNA sample with the repair enzyme cocktail reduced the number of G-to-T variants to baseline levels consistent with 8-oxo-dG damage being the cause of the excess G-to-T variants in the unrepaired samples.

This corresponds to approximately one erroneous call per cancer gene. In summary, our data demonstrate a direct link between damage and the ability to accurately call variants at very low and low to moderate frequency, a frequency typically found for somatic variants.

To assess the extent that damage affects somatic variant calls in cancer samples, we used Varscan (*16*), a popular analysis tool, to identify germline and somatic variants for all TCGA tumor samples with matched tumor-normal pairs (*11*). We estimated the effect of damage on both the high-confidence and total candidate variants identified by Varscan. Before variant calling, R1 and R2 reads were independently grouped to assess the global balance of somatic mutation calls between the two groups. Analogous to GIV, an excess of somatic mutation calls in one group represents erroneous calls caused by damage. A large excess of G-to-T compared with C-to-A somatic variants was found for most data sets (Fig. 4, A and B). Moreover, the fraction of G-to-T variants compared with other variants increased with the estimated damage measured by the GIV_{G_T} score (Fig. 4B). Importantly, data sets of samples predicted to be severely damaged showed an excess of high-confidence G-to-T somatic variants, demonstrating that damage affects high-confidence somatic mutation calls in these samples (Fig. 4C). In contrast, the fraction of G-to-T germline variants was constant across samples and showed no excess in the R1 reads (Fig. 4D), as expected for high-frequency variants.

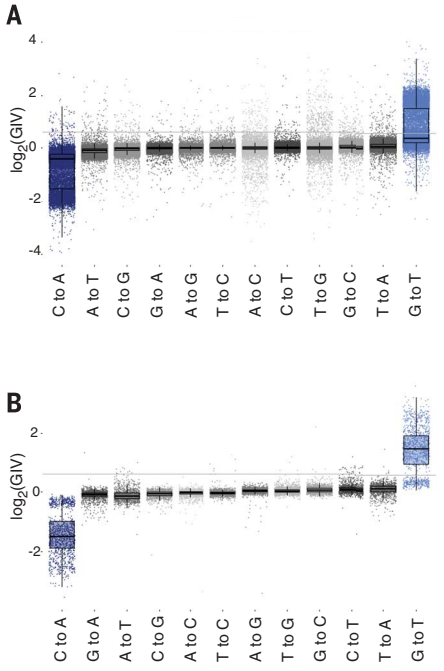


Fig. 2. GIV scores (*y* axis) for the 12 nucleotide substitution classes (*x* axis). (**A**) The 1000 Genomes Project data set. (**B**) A subset of the TCGA data set. Each point represents the GIV score of a single sequencing run downsampled to 5 million reads. The solid gray line denotes a GIV of 1.5. The bimodal distribution of points observed in G-to-T and C-to-A substitution classes corresponds to sequencing runs with damage and without or limited amounts of damage, respectively.

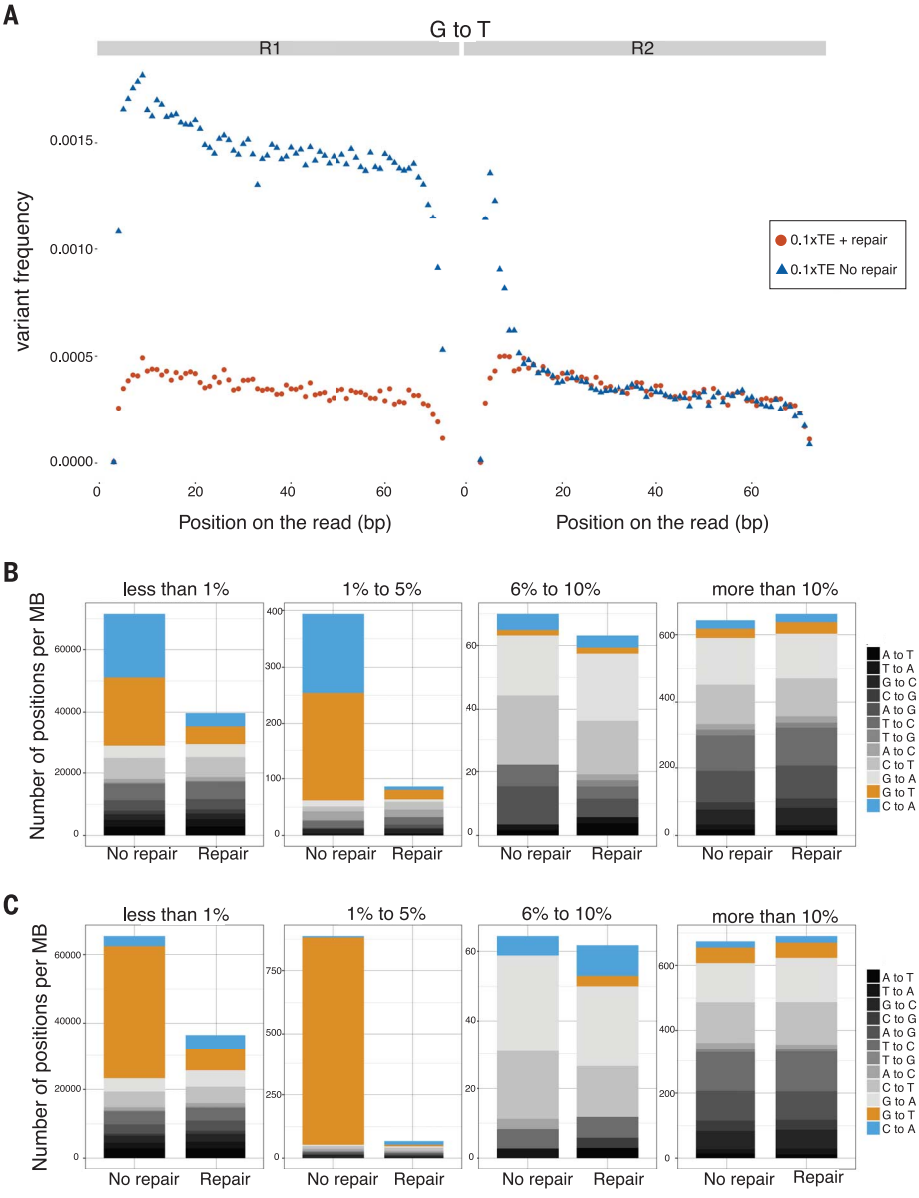
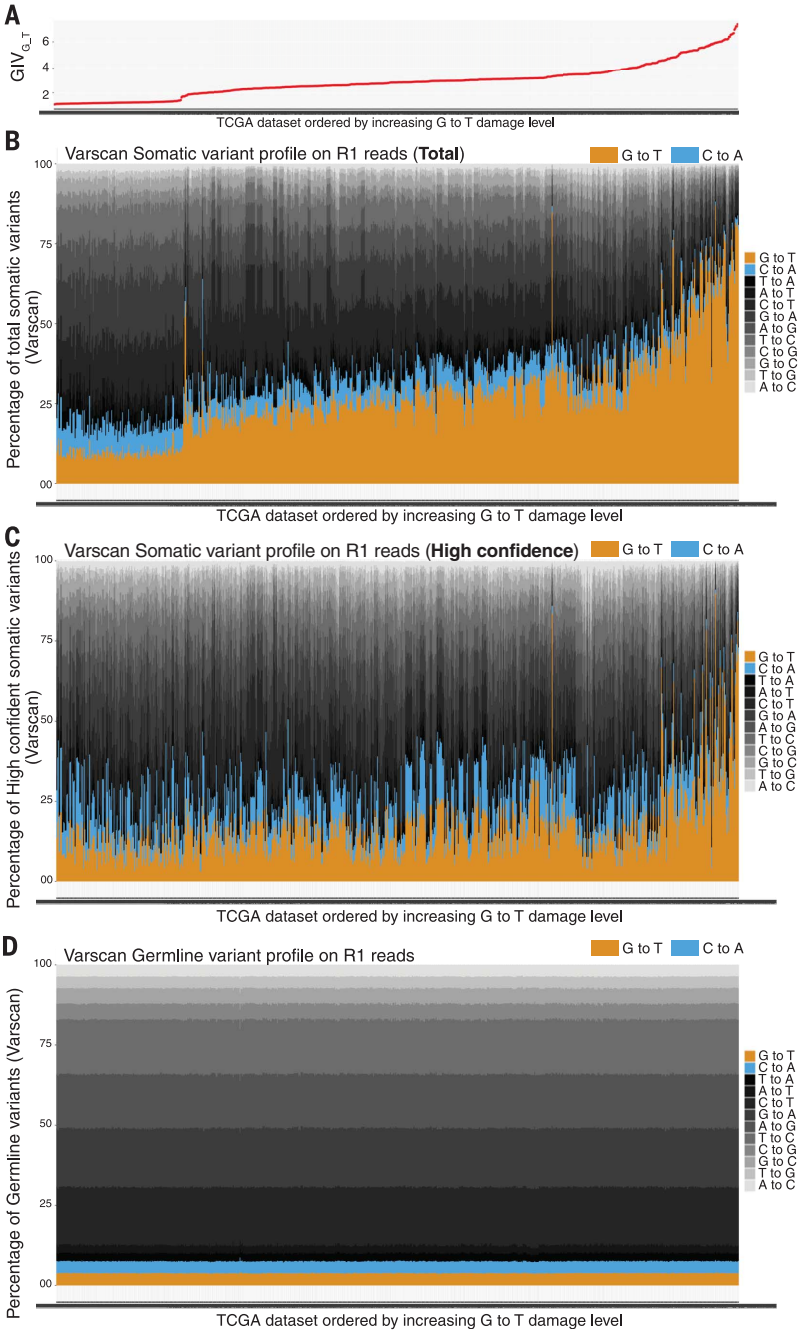


Fig. 3. Target enrichment experiment. (**A**) G-to-T variant profiles across reads R1 and R2 with (red) and without (blue) DNA repair. (**B**) Number of variants per megabase (MB) of sequence per type (orange denotes G-to-T variants, blue denotes C-to-A variants, and shades of gray denote all other variants) at frequencies indicated above the respective graph. (**C**) As in (B), except that only R1 reads were used for variant calling.

Next, we estimated the false-positive rate of somatic variant calls and found that 78% of tumor samples have more than 50% false-positive G-to-T somatic variant calls. Furthermore, the percentage of false positives strongly correlated (r = 0.79) with the estimated damage in tumor samples (Fig. 4E). This correlation between damage and false-positive variant calls indicated that damage is a direct cause of erroneous identification of somatic variants. A smaller subset of the TCGA data set was also identified with a large excess of both total and high-confidence somatic variant calls of the C-to-T type (fig. S7). Together, these results highlight a major confounding effect of damage, including high-confidence somatic mutation calls in the TCGA data sets.

Finally, we evaluated how damage is affecting current TCGA reference variant files. We downloaded the lung adenocarcinoma variant call format files that the TCGA recently generated as part of their annotation workflow and focused on high-confidence variant calls that passed all filters. Focusing on damage leading to G to T, we classified samples as weak or no damage (GIV_{G_T} < 1.5) and heavy damage (GIV_{G_T} > 4.5). The heavy damage group showed an overall moderate increase in the fraction of G-to-T and C-to-A candidate variants for all callers, with Mutect2 (*17*) showing a significant (P < 0.05) difference in distributions (fig. S8). Mutect2 variant profiles displayed a 9% average increase in the fraction of variants being either G-to-T or C-to-A in data sets predicted to be heavily damaged compared with weak or no damage data sets, suggesting that large numbers of variants called with high confidence are derived from artifactual damage. This result is predicted to affect the accurate identification of individual loci and may lead to incorrect diagnostic conclusions in those damaged samples.

To distinguish true from artifactual somatic variants, standard strategies include increasing sequencing coverage, setting stringent variant frequency thresholds, and applying postprocessing computational filters to derive high-confidence variant calls. These stringent criteria can minimize the effect of damage detected genome-wide, as seen for the TCGA variant profiles. Applying stringent criteria, however, does not guarantee the elimination of all errors from damage and, more important, can increase the false-negative rate. For example, variant-calling algorithms can include strand bias to eliminate artifacts, but when faced with limited numbers of variant reads there is a reasonable chance that all evidence reads derived from the same strand orientation, even for genuine variants. Thus, filtering steps are de facto inferior substitutes to preventing



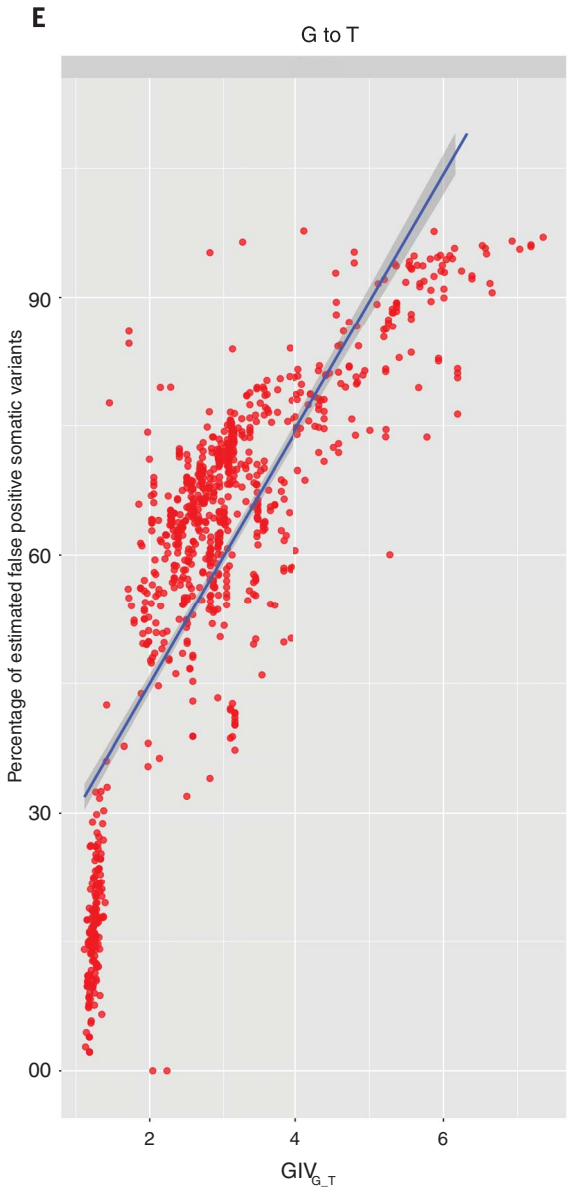


Fig. 4. Variants identified in TCGA data sets. (**A**) About 1800 tumor sequencing runs sorted by increasing GIV_{G_T} score. (**B**) Somatic variant profiles (Varscan) for the tumor samples sorted by increasing GIV_{G_T} scores. The fraction of G-to-T (orange) somatic variant calls is higher than C-to-A (blue) for most data sets, and the fraction of G-to-T calls increases with increasing GIV_{G_T} score. (**C**) As in (B), using the high-confidence somatic variant calls from Varscan. (**D**) As in (B), except the germline variant calls are represented. (**E**) Estimated false-positive rate (in %) (*y* axis) of somatic G-to-T candidate variants found using Varscan as a function of the GIV_{G_T} score (*x* axis).

mutagenic DNA damage from occurring in the first place.

In this work, DNA repair has been used to specifically eliminate oxidative damage in our experimental setup for the purpose of evaluating the GIV score and understanding how damage affects variant calling. Additional work will be required to properly identify conditions that will be effective in eliminating damage from TCGA, 1000 Genomes Project samples, and sequencing samples in general.

REFERENCES AND NOTES

1. I. Martincorena, P. J. Campbell, *Science* **349**, 1483–1489 (2015).
2. L. B. Alexandrov *et al.*, *Nat. Genet.* **47**, 1402–1407 (2015).
3. M. R. Stratton, *Science* **331**, 1553–1558 (2011).
4. D. A. Landau *et al.*, *Cell* **152**, 714–726 (2013).
5. K. Anderson *et al.*, *Nature* **469**, 356–361 (2011).
6. B. Ewing, P. Green, *Genome Res.* **8**, 186–194 (1998).
7. H. Do, A. Dobrovic, *Clin. Chem.* **61**, 64–71 (2015).
8. A. W. Briggs *et al.*, *Nucleic Acids Res.* **38**, e87 (2010).
9. A. M. Newman *et al.*, *Nat. Biotechnol.* **34**, 547–555 (2016).
10. M. Costello *et al.*, *Nucleic Acids Res.* **41**, e67 (2013).
11. Materials and methods are available as supplementary materials.
12. K. C. Cheng, D. S. Cahill, H. Kasai, S. Nishimura, L. A. Loeb, *J. Biol. Chem.* **267**, 166–172 (1992).
13. J. Tchou, A. P. Grollman, *Mutat. Res.* **299**, 277–287 (1993).
14. A. Auton *et al.*, *Nature* **526**, 68–74 (2015).
15. J. Bendl *et al.*, *PLOS Comput. Biol.* **12**, e1004962 (2016).
16. D. C. Koboldt *et al.*, *Genome Res.* **22**, 568–576 (2012).
17. K. Cibulskis *et al.*, *Nat. Biotechnol.* **31**, 213–219 (2013).

ACKNOWLEDGMENTS

We would like to thank S. Kaiser, A. Messelaar, T. Vincze, and C. Lin for information technology and compliance with the TCGA hosting guidelines; L. Mazzola, J. Bybee, and D. Rivizzigno for sequencing; and R. Roberts, W. Jack, T. Carlow, A. Gardner, H. Runz, E. Dimalanta, and S. Russello for critical comments. The results shown here are in part based on data generated by the 1000 Genomes Projects and the TCGA Research Network: https://cancergenome.nih.gov. This research was supported by New England Biolabs Inc. L.E., T.C.E., L.C., and P.L. are 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. Sequencing data have been deposited at the European Nucleotide Archive under accession number PRJEB16681. The algorithm for the GIV score is available at https://github.com/Ettwiller/Damage-estimator. Disclosure declaration: All authors are employees of New England Biolabs Inc. Ethics statement: DNA samples in this study were collected under the BioChain Institute Inc. Institutional Review Board (IRB). This IRB is registered with the Office for Human Research Protections (OHRP), registration number IRB00008283, and has been issued with Federal Wide Assurance (FWA), FWA00017355, for the Protection of Human Subjects for Institutions within the United States by OHRP of the U.S. Department of Health and Human Services. The DNA was collected using an informed consent form approved by Biochain's IRB. The nature and possible consequences of the studies were explained in the informed consent form.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/355/6326/752/suppl/DC1
Materials and Methods
Supplementary Text 1 to 5
Figs. S1 to S8
Tables S1 and S2
Reference (*18*)

23 August 2016; accepted 23 January 2017
10.1126/science.aai8690

Editor's Summary

When is a mutation a true genetic variant?

Large-scale sequencing studies have set out to determine the low-frequency pathogenic genetic variants in individuals and populations. However, Chen *et al.* demonstrate that many so-called low-frequency genetic variants in large public databases may be due to DNA damage. They scored libraries sequenced with and without a DNA damage–repairing enzymatic mix to assess the proportion of true rare variants. It remains to be seen how best to repair DNA before sequencing to provide more accurate assessments of mutation.

Science, this issue p. 752