Study shows higher than expected sequencing errors in public databases

17 February 2017, by Bob Yirka

A depiction of the double helical structure of DNA. Its four coding units (A, T, C, G) are color-coded in pink, orange, purple and yellow. Credit: NHGRI

(Medical Xpress)—A team of researchers with New England Biolabs Inc. (NEB) has found that sequenced DNA samples held in public databases had higher than expected low-frequency mutation error rates. In their paper published in the journal *Science*, the team describes how they created an algorithm that is able to calculate an error rate for samples in a database and what it showed when run on two public genome databases.

Researchers involved in studying the role DNA plays in cell mutations that lead to cancerous tumors rely on the accuracy of databases that hold sequencing information—those looking for commonalities, for example, among different groups of people rely on information in such databases when attempting to isolate trends. Such studies involve comparing the genomes of different people with low-frequency mutations versus the general population and using what they find to build cancer datasets. But now, the accuracy of public databases has been called into question by work done by the team at NEB, which in turn calls into question the accuracy of the cancer datasets.

To measure the accuracy rate of a given dataset, the researchers created an algorithm that could be used to count the numbers of sequences showing mutations due to damage during the sequencing process versus those that happened naturally. The team then used their algorithm to calculate error rates for several public databases—most notably the 1000 Genomes Project and part of the TCGA database—they report that they found error rates of 41 percent and 73 percent respectively.

The researchers note that their algorithm is not capable of revealing the source of unnatural damage, but suggest it is likely due to certain sample preparation techniques used prior to sequencing. They also point out that other algorithms have been developed for sequencers to test their own work for errors, but due to lack of a compelling reason, they have not been widely used. They suggest DNA sequencers begin doing so. They also note that new tools have been developed that could help minimize DNA damage during preparation and that their use could improve the accuracy of public databases.

More information: Lixin Chen et al. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification, *Science* (2017). DOI: 10.1126/science.aai8690

Abstract

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.

APA citation: Study shows higher than expected sequencing errors in public databases (2017, February 17) retrieved 18 November 2019 from https://medicalxpress.com/news/2017-02-higher-sequencing-errors-databases.html