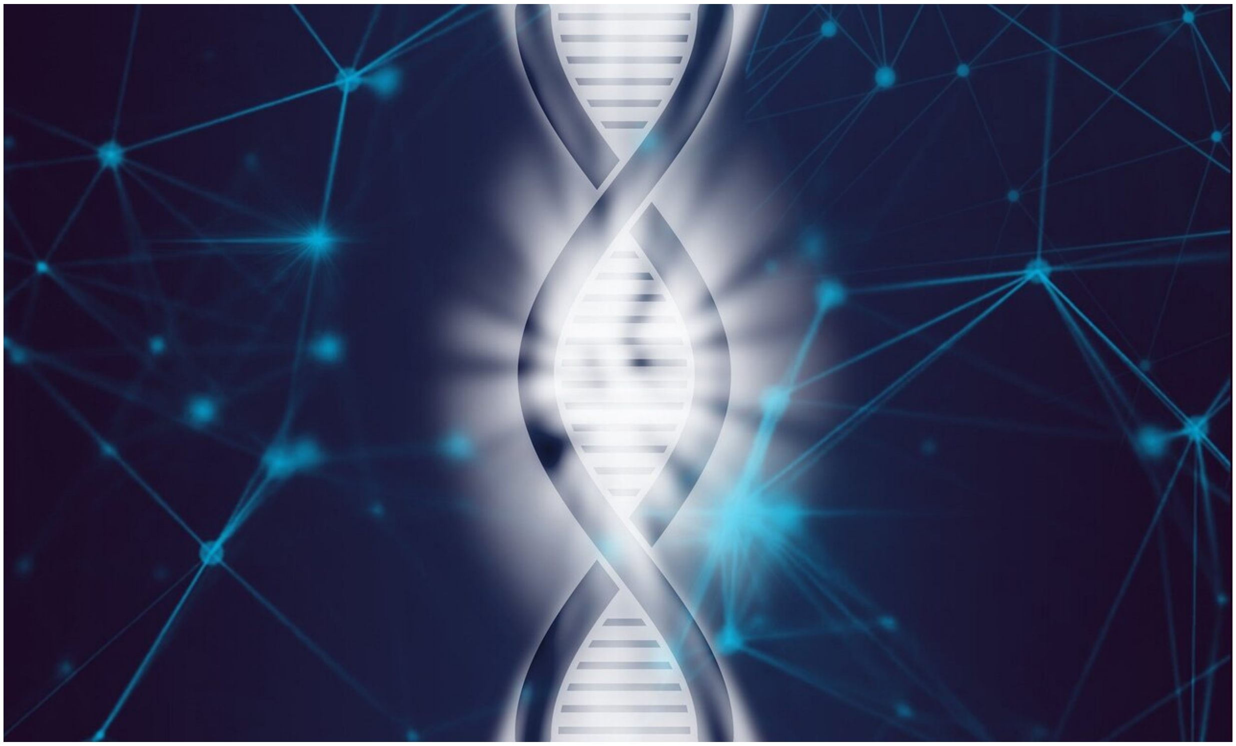
Study finds widespread misinterpretation of gene expression data

December 18 2019



Credit: CC0 Public Domain

Reproducibility of research data is a major challenge in experimental biology. As data generated by genomic-scale techniques increases in complexity, this concern is becoming more and more worrisome.

RNA-seq, one of the most widely used methods in modern molecular biology, allows in a single test the simultaneous measurement of the expression level of all genes in a given sample. New research by a Tel Aviv University group identifies a frequent technical bias in data generated by RNA-seq technology, which often leads to false results.

The study was conducted by Dr. Shir Mandelbaum, Dr. Zohar Manber, Dr. Orna Elroy-Stein and Dr. Ran Elkon at TAU's Sackler Faculty of Medicine and George S. Wise Faculty of Life Sciences and was published on November 12 in *PLOS Biology*.

"Recent years have witnessed a growing alarm about false results in biological research, sometimes referred to as the reproducibility crisis," Dr. Elkon, lead author of the study, says. "This study emphasizes the importance of proper statistical handling of data to lessen the number of misleading findings."

A main goal of RNA-seq experiments is to characterize biological processes that are activated or repressed in response to different conditions. The researchers analyzed dozens of publicly available RNA-seq datasets to profile the cellular responses to numerous stresses.

During the research, the scientists noticed that sets of particularly short or long genes repeatedly showed changes in the expression level measured by the apparent number of RNA transcripts from a given gene. Puzzled by this recurring pattern, the team wondered whether it reflected some universal biological response common to different triggers or whether it stemmed from some experimental condition.

To tackle this question, they compared replicated samples from the same biological condition. Differences in gene expression between replicates can reflect technical effects that are not related to the experiment's biological factor of interest. Unexpectedly, the same pattern of particularly short or long genes showing changes in expression level was observed in these comparisons between replicates. This pattern is the result of a technical bias that seemed to be coupled with gene length, the researchers say.

Importantly, the TAU researchers were able to show how the length bias they detected in many RNA-seq datasets led to the false identification of specific biological functions as cellular responses to the conditions tested.

"Such misinterpretation of the data could lead to completely misleading conclusions," Dr. Elkon concludes. "In practical terms, the study also shows how this bias can be removed from the data, thus filtering out false results while preserving the biologically relevant ones."

More information: Shir Mandelboum et al, Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias, *PLOS Biology* (2019). DOI: 10.1371/journal.pbio.3000481

Provided by Tel Aviv University

Citation: Study finds widespread misinterpretation of gene expression data (2019, December 18) retrieved 3 September 2024 from https://phys.org/news/2019-12-widespread-misinterpretation-gene.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.