MedicalXpress

Researchers discover extensive mislabeling of gene expression samples

October 12 2016

At least 1 in 3 gene expression studies contain mislabelled samples, according to a new study published in F1000Research. As correct identification of the samples is central to data analysis, studies based on data from mislabelled samples could reach incorrect conclusions.

After discovering labelling errors while reanalysing gene expression datasets from a Parkinson's disease study, Lilah Toker, Min Feng and Paul Pavlidis of the University of British Columbia decided to investigate how often these mistakes occur. Their findings, published in the F1000Research channel Preclinical Reproducibility & Robustness, have now passed peer review.

The researchers used an elegant approach to detect whether a sample was mislabelled by assessing the expression levels of genes on the sex-specific chromosomes: Females specifically express some genes located on the X chromosome, while only males express genes located on the Y chromosome. Expression level of the sex-specific genes can be compared to the sex stated on the sample's label to determine if mislabelling has occurred.

Of 70 human tissue datasets studied, comprising 4,043 samples in total, the team found that 46% datasets contained at least one discrepancy between the sample's sex-specific gene expression levels and the sex written on the label. Based on this data they calculate that at least 33%, and up to 60%, of all gene expression studies contain mislabelled samples. The authors note this might be only a snapshot of a wider

problem, as their method cannot detect cases where the mislabelling did not affect the sex of the sample - such as a wrong tissue sample or a mix-up of two samples from two subjects with the same sex).

The authors explored at which stage of data collection, analysis or report the mislabelling occurred. While in majority of the cases the exact origin could not be identified, the authors found that mislabellings were often already present at the stage of data analysis, and in several cases the mislabelling was traced back to laboratory test tube mix-ups.

Lilah Toker said: "Researchers have long been aware of the value of sex markers for quality control, so it was surprising to find such obvious problems in so many studies. We hope our study encourages greater diligence."

Leonard P. Freedman of the Global Biological Standards Institute in Washington DC, who openly reviewed the paper, said: "This is an excellent paper highlighting the importance of sample annotation as a critical contributor to reproducible research."

Hans van Bokhoven of the Radboud University Medical Center, who also approved the article as a reviewer, said: "While these figures are already alarming, the actual number of mismatches is likely to be higher, because the gender-analysis can only identify discrepancies based on a gender-mismatch and will not detect mislabelling of samples of the same gender and case-control samples."

As inaccurate labelling has the potential to seriously undermine the validity and reuse of gene expression data, the authors argue that such sex-specific gene expression checks should become routine.

More information: Lilah Toker et al, Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies,

F1000Research (2016). DOI: 10.12688/f1000research.9471.2