www.nature.com /articles/d41586-021-02211-4

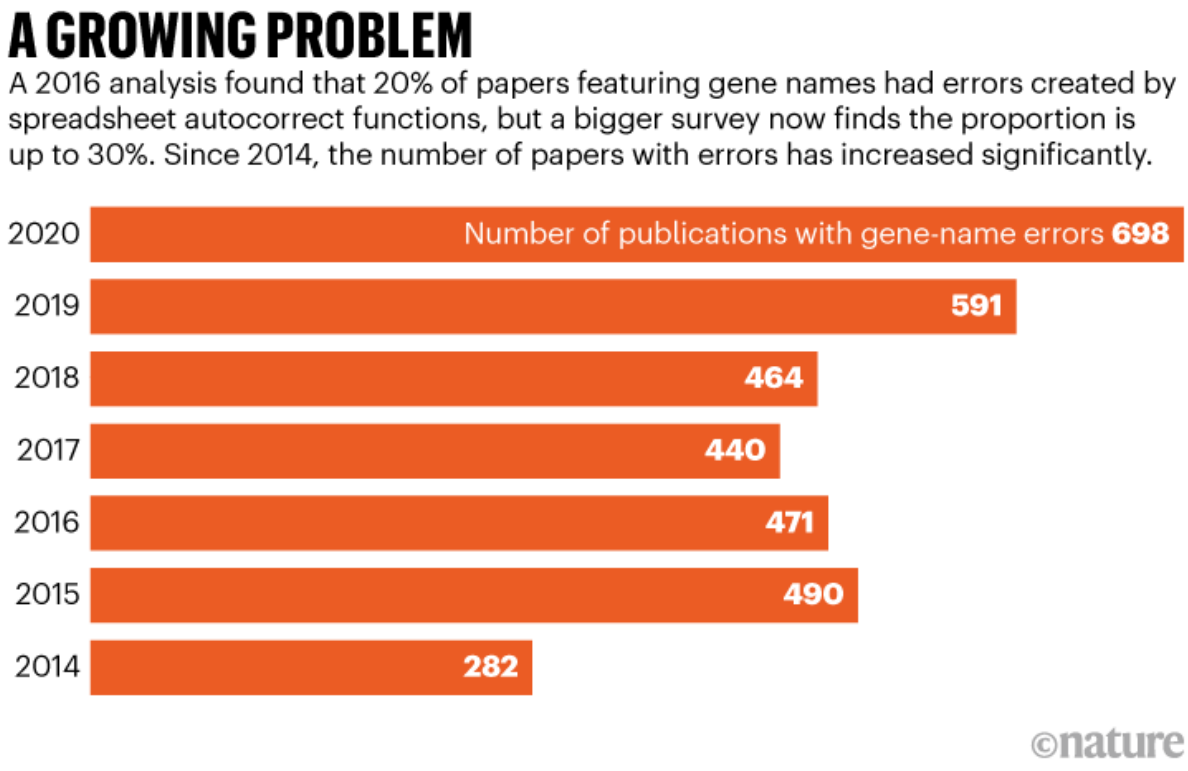
Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

Embarrassing autocorrect mistakes are common fodder for Internet listicles and Twitter threads. But they are also the bane of geneticists using spreadsheet programs such as Microsoft Excel. Five years after a study showed that autocorrect problems were widespread, the academic literature is still littered with error-riddled spreadsheets, according to an analysis of published gene lists. And the problem may be even worse than previously realized.

The long-standing issue often occurs when the abbreviated form of a gene's name — known as a gene symbol — is incorrectly recognized as a date and autocorrected as such by Excel or Google Sheets. For example, *SEPT4* (septin 4) and *MARCH1* (membrane associated ring-CH-type finger 1) will be automatically changed to *4-Sep* and *1-Mar*.

"It can have a significant impact on your research," says molecular biologist Auriol Purdie at the University of Sydney in Australia. Having worked with gene-microarray and gene-transcription data sets for two decades, Purdie is familiar with the inadvertent errors. But she says the problem frequently catches out beginners.



Distorting results

Purdie works to identify gene networks involved in the early stages of disease in sheep and cattle. If a spreadsheet alters the gene names, those genes are lost when the data are imported into gene-network-analysis software, and this can distort results. The program “will tell you you've lost a bunch of your genes”, she says, but won’t indicate which ones. And when dealing with data sets that contain 20,000 genes, manually comparing lists to identify which genes have been lost is an onerous task, Purdie adds.

The problem was first documented in 2004, when Barry Zeeberg, a molecular pharmacologist at the National Cancer Institute in Bethesda, Maryland, and his colleagues warned of changes to gene symbols when processing genomics data[1].

In 2016, Mark Ziemann and his colleagues at the Baker IDI Heart and Diabetes Institute in Melbourne, Australia, quantified the problem. They found that one-fifth of papers in top genomics journals contained gene-name conversion errors in Excel spreadsheets published as supplementary data[2]. These data sets are frequently accessed and used by other geneticists, so errors can be perpetuated and distort further analyses.

However, despite the issue being brought to the attention of researchers — and steps being taken to fix it — the problem is still rife, according to an updated and larger analysis led by Ziemann, now at Deakin University in Geelong, Australia[3]. His team found that almost one-third of more than 11,000 articles with supplementary Excel gene lists published between 2014 and 2020 contained gene-name errors (see ‘A growing problem’).

Simple checks can detect autocorrect errors, says Ziemann, who researches computational reproducibility in genetics. But without those checks, the errors can easily go unnoticed because of the volume of data in spreadsheets.

Changes to naming conventions

In 2017, the HUGO Gene Nomenclature Committee (HGNC) — which standardizes human-gene names — announced that it would take the drastic measure of changing the gene symbols for commonly affected genes, because community-outreach efforts (including a 2016 video on YouTube) had failed to solve the problem. Since then, 27 gene symbols have been updated, including *SEPT4* (now *SEPTIN4*) and *MARCH1* (now *MARCHF1*).

The move was a departure from the committee’s preference for keeping names stable, says Elspeth Bruford, who coordinates the HGNC from the European Bioinformatics Institute in Hinxton, UK. Last year, the committee published guidelines to reflect the new rule for modifying gene symbols in cases where data handling is affected[4]. Other gene-naming bodies have followed suit.

But it might be too soon to see any change to the frequency of errors in the literature, says Bruford, because published data sets often contain outdated gene lists. “It’s going to take years for this to percolate through,” she says, which is why the HGNC recommends that researchers access the most recent data from public databases, and that journals request authors to do so before publication.

Since the beginning of the year, Ziemann has published a monthly leader board of offending journals, which frequently features well-known titles such as *Nature Communications*, *eLife*, *PLoS Genetics* and

Scientific Reports. Ziemann says this is probably because articles published in these journals contain more gene lists and larger data sets.

Avoid or adapt

One solution is to avoid using spreadsheets, he suggests. Although some — such as the open-source programs LibreOffice and Gnumeric — don't have the problem, spreadsheets are hard to audit. "If there's a problem, it's not readily apparent where the problem happened," because there's no record of what steps the software took, he says.

Some computational biologists use scripted computer languages, such as Python and R. These don't autocorrect gene symbols, says Ziemann, and researchers can trace the source of errors. However, they require users to learn the computer language so that they can write code to analyse data.

That's something Purdie says she doesn't have time for. She has adapted to Excel's quirks, adding apostrophes before commonly affected genes to prevent the conversion, or pre-formatting spreadsheet cells before importing data. "It's one of those things that I just accept," she says.

Bruford says the autocorrect issue in Excel is unlikely to be fixed any time soon. "We're a small user base, compared to all the users of Excel," she says, and Microsoft has never indicated that it will alter its software to accommodate the genetics community.

For those persisting with problematic software, Ziemann recommends a quick check before sharing or publishing data. Sorting data by gene symbol can bring date-conversion errors to the top, he says.