nature genetics

Legible ledgers

A prevalent but trivial systematic error in supplementary tables provides a reminder that genomic and other large data files are most usable when they are readable by both humans and machines. It is best practice to deposit large files in public databases and to provide accession links for peer review rather than to delay data deposition until publication.

Preparation of tables of gene names for manual inspection or publication can entail exporting or copying data from bioinformatics workflows into more widely used office software. Since 2004, this practice has been known to cause irreversible reformatting of a small number of gene names in large tables (*BMC Bioinformatics* **5**, 80, 2004). If authors are unaware of the default settings of the useful Microsoft Excel spreadsheet program, gene names such as *SEPT11* can automatically be converted to a date such as 11-Sep and RIKEN identifiers can be converted to floating point numbers. Within a list of tens of thousands of gene names, these changes can escape proofreading by human eyes. A more recent survey (*Genome Biol.* **17**, 177, 2016) claims that a large number of published articles are still affected by this problem. Not all articles have supplementary spreadsheets, but we examined 358 tables in Excel format appended to 83 articles published in this journal in the first nine months of this year. Of these, supplementary tables from 15 articles (18%) were affected by the altered formatting of gene identifiers. We found that 24 of 358 supplementary tables (7%) and 1,925 of 1.6 million gene names (0.12%) were incorrect.

It is rare that these small changes would have any consequence to the scientific conclusions of the published work or their reproducibility and reuse, but they occur often enough to surprise and irritate some analysts, particularly in interdisciplinary research. We have posted a list of supplementary tables that we know to contain some incorrect gene names to the journal blog (http://blogs.nature.com/freeassociation/) for the convenience of our readers without intending any criticism of the software, the publications or the authors, as we have decided that these problems are insufficient to justify either formal published correction or replacement of the files. Human readers of supplementary tables who are alert to these sporadic substitutions can identify at least 37 commonly affected loci (*FEB1–FEB11*, *MARCH1–MARCH11*, *SEPT1–SEPT14* and *DEC1*) using a list of correct gene names (http://www.genenames.org/) and redundant information such as genome coordinates from the tables.

There are, however, some valuable lessons within the big picture of increasing the reliance of research upon computed analyses in combination with expert scrutiny. The first of these is of course for experts to distinguish easily made and easily detected misprints from important scientific errors and to explain the difference to non-specialists. The second point is to help people read and machines to work by providing properly formatted data. Genome coordinates relative to a stated reference sequence are more generally useful than gene symbols, as in many tables the number of anonymous probes and loci with regional coordinates exceeds the number of gene names. Unfortunately, most supplementary tables are not machine readable, often because the tables do not follow the convention that "each cell of a table at the intersection of a row and a column contain a single entity" (*Nat. Genet.* **43**, 1, 2011), Even more frequently, supplementary tables are difficult to operate with a script because cells are sometimes merged for table titles and legends, leaving the table with an unpredictable number of columns in each row. This also temporarily hinders identification of transformed gene names via alphabetical sorting.

To anticipate and prevent future problems, we recommend that all authors check their supplementary tables before submission for common errors and human readability and submit machine-readable versions of large supplementary tables and any associated scripts to one of the recommended data repositories (http://www.nature.com/sdata/policies/repositories), including live accession codes in the submitted manuscript to make the peer referees' job easier. ■