chem-bla-ics

Excel messes up your data analysis :)

Published August 1, 2007

Citation
Willighagen, E. (2007, August 1). Excel messes up your data analysis :). *Chem-bla-ics*. https://doi.org/10.59350/81926-4bz44

Keywords
Bioinfo, Excel

Abstract
Well, no wonder: Excel is meant to be used to process money flows. Anyway, greyarea pointed me to this nice blog item from March 2006. It discusses a 2004 article in BMC Bioinformatics Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics by Barry Zeeberg et al. (DOI:10.1186/1471-2105-5-80). Hence, the importance of semantics and proper markup languages.

Well, no wonder: Excel is meant to be used to process money flows. Anyway, greyarea pointed me to this nice blog item from March 2006. It discusses a 2004 article in BMC Bioinformatics *Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics* by Barry Zeeberg et al. (DOI:10.1186/1471-2105-5-80). Hence, the importance of semantics and proper markup languages. The quotes are illustrative:

> *When we were beta-testing [two new bioinformatics programs] on microarray data, a frustrating problem occurred repeatedly: Some gene names kept bouncing back as "unknown." A little detective work revealed the reason: ... A default date conversion feature in Excel ... was altering gene names that it considered to look like dates. For example, the tumor suppressor DEC1 [Deleted in Esophageal Cancer 1] was being converted to '1-DEC.' Figure 1 lists 30 gene names that suffer an analogous fate.*
>
> ...
>
> *There is another default conversion problem for RIKEN clone identifiers identifiers of the form nnnnnnnEnn, where n denotes a digit. These identifiers are comprised of the serial number of the plate that contains the library, information on plate status, and the address of the clone. A search ... identified more than 2,000 such identifiers out of a total set of 60,770. For example, the RIKEN identifier "2310009E13" was converted irreversibly to the floating-point number "2.31E+13." A non-expert user might well fail to notice that approximately 3% of the identifiers on a microarray with tens of thousands of genes had been converted to an incorrect form, yet the potential for 2,000 identifiers to be transmogrified without notice is a considerable concern. Most important, these conversions to an internal date representation or floating-point number format are irreversible; the original gene name cannot be recovered.*

Is this the article that made all bioinformaticians turn to R?