

Literature Lab™ Gene Thesaurus excels in eliminating gene name errors proven to be widespread in the scientific literature

September 15, 2016 – for immediate release

Contact: Damon Anderson, PhD – danderson@acumenta.com - www.acumenta.com

A recent comment in Genome Biology (Ziemann et al. Genome Biology (2016) 17:177) revisited the issue of the Excel software (under default settings) mistakenly converting gene symbols to dates and floating-point numbers, as originally described in 2004 (Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics. 2004. 5:80.)

The results of the authors' analysis of 18 journals and 7467 gene lists attached to 3597 published papers from the period 2005 to 2015, confirmed gene name errors in 987 supplementary files from 704 published articles. Of the selected journals, the proportion of published articles with Excel files containing gene lists that are affected by gene name errors is 19.6 %. This striking percentage among a range of high impact journals has sent shock waves through the genome research community and has received massive media attention. At the heart of the matter is the apparent persistence of gene name errors and inadvertent gene symbol conversions that continue to plague supplementary gene lists, an important resource to the genomics community.

The Acumenta Gene Thesaurus is a repository of human gene and protein nomenclature. It contains symbols, names and aliases gathered from major genomic data repositories combined with extensive human and machine-assisted curation to assure comprehensive and precise searching of literature on gene and protein topics. The Gene Thesaurus is integrated into the Literature Lab™ gene list import module, assuring accurate and efficient resolution of nomenclature issues. The import system traps all records that have been inadvertently modified by Excel.

Once gene symbols are flagged, corrected and validated, the gene list then enters the statistical analysis engine that drives the identification of significant associations between gene lists and key biological and biochemical concepts in the published literature. The fundamental process and abilities of Literature Lab™ are unique and empowering among other functional analysis platforms.

As a demonstration of the 'quality control' attributes of the Literature Lab™ Gene Thesaurus, an excel file loaded with MARCHx and SEPTx genes was submitted for validation in the initial step of Literature Lab™ analysis. Consider a more likely real world scenario where a few of these genes may be buried in a list of dozens or hundreds of genes. The following is the Excel-processed data as viewed in the Editor after being validated:

