## Literature Lab<sup>™</sup> 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<sup>™</sup> 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<sup>™</sup> are unique and empowering among other functional analysis platforms.

As a demonstration of the 'quality control' attributes of the Literature Lab<sup>™</sup> Gene Thesaurus, an excel file loaded with MARCHx and SEPTx genes was submitted for validation in the initial step of Literature Lab<sup>™</sup> 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 Excelprocessed data as viewed in the Editor after being validated:

	Gene List Litle:						
Sho	ort Series Label:		All Genes Down Regulated?				
0	Driginal ID Type: Symb	ool(Human) 👻	Use Gene Thesaurus	Clear the G	iene List	Import a Gene List	
	Enter you	Ir gene ids in the first column or use the "Import a	ene List" button. Click on a column header	to sort.	_	Define Columns	
Orig ID	Symbol	Name	St	atus			
G	A 1BG	alpha-1-B glycoprotein	)	/alid			
F	A1CF	APOBEC1 complementation factor		/alid			
20	MAPI	microtubule associated protein tau	Line	/alid			-
30	2	2	Linic	lentified			
30	2	2	Unic	lentified			-
31			Unic	lentified			
34		2	Unic	lentified			
35	2	2	Unid	lentified			
36	2	8	Unic	lentified	-		
137			Unid	lentified			
RCKS	MARCKS	myristoylated alanine rich protein kinase C substrate		/alid	1		
RCKSL1	MARCKSL1	MARCKS-like 1		/alid	1		
SECS	SEPSECS	Sep (O-phosphoserine) tRNA:Sec (selenocysteine) tRNA	synthase	/alid			
14			Unic	entified			
523			Unic	entified			
24			Unic	entified			
25	-		Unic	lentified			
515	2	2	Unic	lentified			
516	2		Linic	lentified			
517	2	20	Unic	lentified	1		
518	8	12	Unid	lentified			
519	0	2	Unid	lentified			
520	2	8	Unid	lentified			-
PT7P1	SEPT7P1	septin 7 pseudogene 1		/alid	5		
521	2		Unic	lentified	1		
522	2		Unic	lentified			
PW1	SEPW1	selenoprotein W, 1		/alid	-		
<	ZYX	zyxin	2	/alid			
		5 0					
		2					
		2					
		2					
		2					
		8					
		1					

The numbers are the Excel numerical values for the dates (for the current year). These do not mesh with the Gene Thesaurus and thus cannot be validated, necessitating the user to quality check the spreadsheet data and make any necessary corrections prior to analysis. The Literature Lab<sup>™</sup> Gene Thesaurus is a fail-safe against inadvertent creation and propagation of erroneous gene symbols. The Thesaurus is a comprehensive and vital component of Literature Lab<sup>™</sup> and eliminates the need for external quality control measures. The result is fast and efficient data processing and comprehensive and precise gene list analysis, thereby leading researchers to think about the output and not worry about the input.

--

Acumenta Biotech, based in Westminster, Massachusetts, is a software and information company serving biotechnology and pharmaceutical firms and life sciences non-profit educational and research institutions. Acumenta Biotech builds applications that automate information gathering, analysis and management tasks on public Web-based and proprietary external and internal databases. Acumenta Biotech's software products and its publication record-derived biological knowledge database improve both the efficiency and quality of research in the life sciences.