BMC Bioinformatics

BioMed Central

Correspondence

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg[†1], Joseph Riss[†2], David W Kane[3], Kimberly J Bussey[1], Edward Uchio[4], W Marston Linehan[4], J Carl Barrett[2] and John N Weinstein[*1]

Address: [1]Genomics & Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research (CCR), National Cancer Institute (NCI), National Institutes of Health (NIH), Bldg 37 Rm 5041, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, [2]Laboratory of Biosystems and Cancer, CCR, Bldg 37 Rm 5032, NIH, 9000 Rockville Pike, Bethesda, MD 20892 USA, [3]SRA International, 4300 Fair Lakes CT, Fairfax, VA 22033 USA and [4]Urologic Oncology Branch, Bldg 10 Rm 2B47, National Institutes of Health, Bethesda, MD 20892 USA

Email: Barry R Zeeberg - barry@discover.nci.nih.gov; Joseph Riss - rissj@helix.nih.gov; David W Kane - david_kane@sra.com; Kimberly J Bussey - busseyk@mail.nih.gov; Edward Uchio - eu8v@nih.gov; W Marston Linehan - linehanm@mail.nih.gov; J Carl Barrett - barrett@mail.nih.gov; John N Weinstein* - weinstein@dtpvx2.ncifcrf.gov

* Corresponding author †Equal contributors

Published: 23 June 2004

BMC Bioinformatics 2004, **5**:80 doi:10.1186/1471-2105-5-80

Received: 05 March 2004
Accepted: 23 June 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/80

Abstract

Background: When processing microarray data sets, we recently noticed that some gene names were being changed inadvertently to non-gene names.

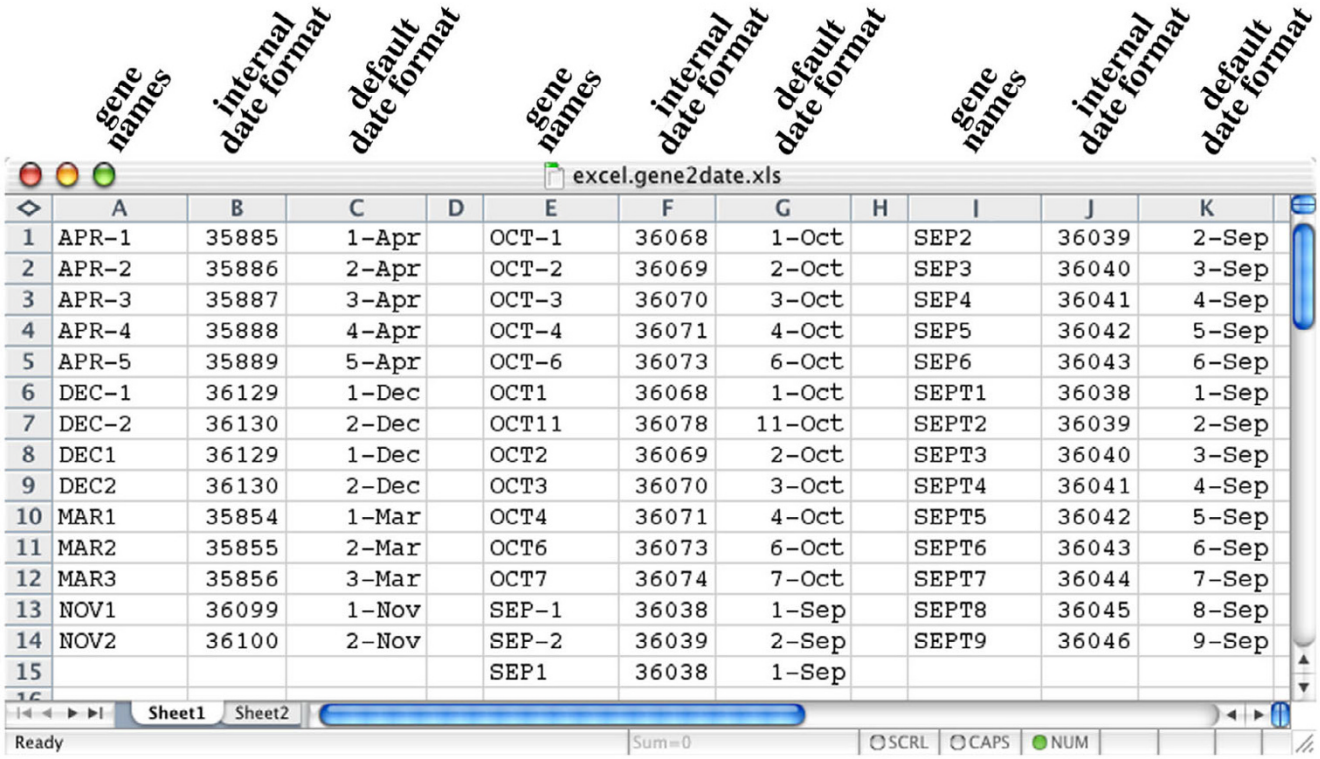
Results: A little detective work traced the problem to default date format conversions and floating-point format conversions in the very useful Excel program package. The date conversions affect at least 30 gene names; the floating-point conversions affect at least 2,000 if Riken identifiers are included. These conversions are irreversible; the original gene names cannot be recovered.

Conclusions: Users of Excel for analyses involving gene names should be aware of this problem, which can cause genes, including medically important ones, to be lost from view and which has contaminated even carefully curated public databases. We provide work-arounds and scripts for circumventing the problem.

Text

MatchMiner [1] and GoMiner [2] are two bioinformatics program packages we published recently in another Biomed Central Journal, Genome Biology. When we were beta-testing those programs on microarray data, a frustrating problem occurred repeatedly: Some gene names kept bouncing back as "unknown." A little detective work revealed the reason: Use of one of the research community's most valuable and extensively applied tools for manipulation of genomic data can introduce erroneous names. A default date conversion feature in Excel (Microsoft Corp., Redmond, WA) was altering gene names that it considered to look like dates. For example, the tumor suppressor DEC1 [Deleted in Esophageal Cancer 1] [3] 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 [4] clone 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 [5]. A search (using the "DNA sequence length search" functionality at http://fantom2.gsc.riken.go.jp/db/search/) identified more than 2,000 such identifiers out of a total set of 60,770. For example, the RIKEN identifier "2310009E13"



	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

Figure 1
Screen shot of Microsoft Excel spreadsheet illustrating errors caused by default conversion of gene names to dates. Columns A, E, and I contain the correct gene names. Columns B, F, and J contain the corresponding underlying internal Excel date representation resulting from the forced default date conversion. Columns C, G, and K contain the corresponding default format date conversions. To create this table, we prepared a tab-delimited text file in which each gene name was repeated three times side by side. The correct gene names in columns A, E, and I were retained by opening this text file with Excel, and selecting "text" mode for columns A, E, and I in the Text Import Wizard Step 3 of 3 that appears while opening a file in Excel. Subsequently, the format menu "number" option (with zero decimal places) was applied to columns B, F, and J to display the internal date format.

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. If one were dealing manually with small numbers of genes, these problems could be detected and then corrected by the tedious, convoluted process described in the legend of Figure 1. But with microarray or other high-throughput data, human proofreading and manual curation are impractical.

The floating-point conversion is not restricted to RIKEN clone identifiers but will affect any clone designation derived from plate coordinates. Although the standard convention is to designate clones officially by library-plate-row-column identification, it is common practice to omit the library reference, particularly if all of the clones come from a single library. Without the library reference in the identifier, all clones from row E of any plate are converted to floating point numbers by Excel. If the library designation is numeric, as it is for RIKEN clones [5], then including it does not solve the problem. Since 96-well plates contain 8 rows and 12 columns, row E represents 12/96 or 12.5% of the clones on the plate; similarly, 6.25% of clones from 384-well plates would be affected. Most libraries contain hundreds of plates, each of which would be subject to this problem.

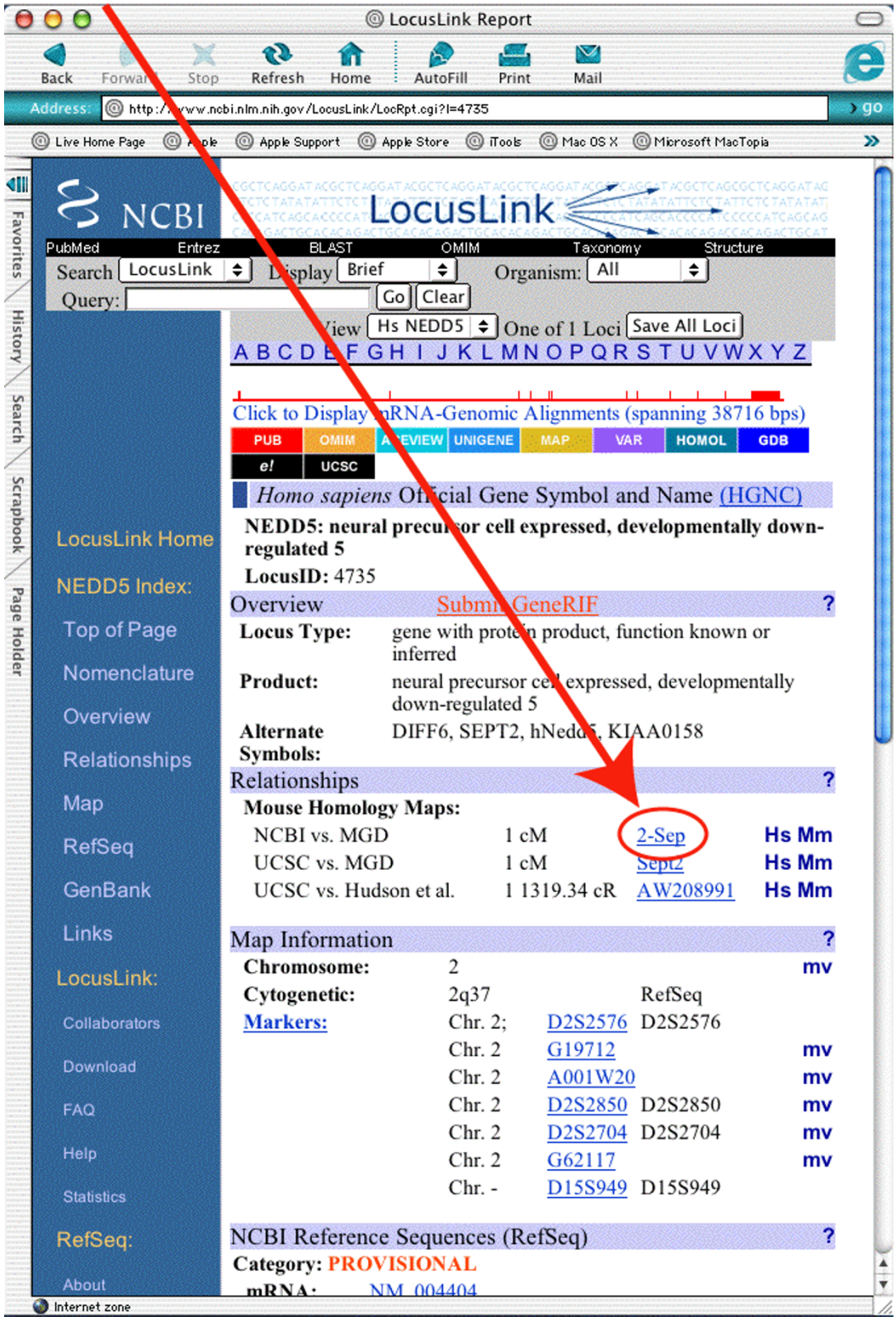


Figure 2
Screen shot of LocusLink from November 12, 2002 illustrating an error caused by default conversion of a gene name to date that had propagated from the human-mouse homology map data (Figure 3).

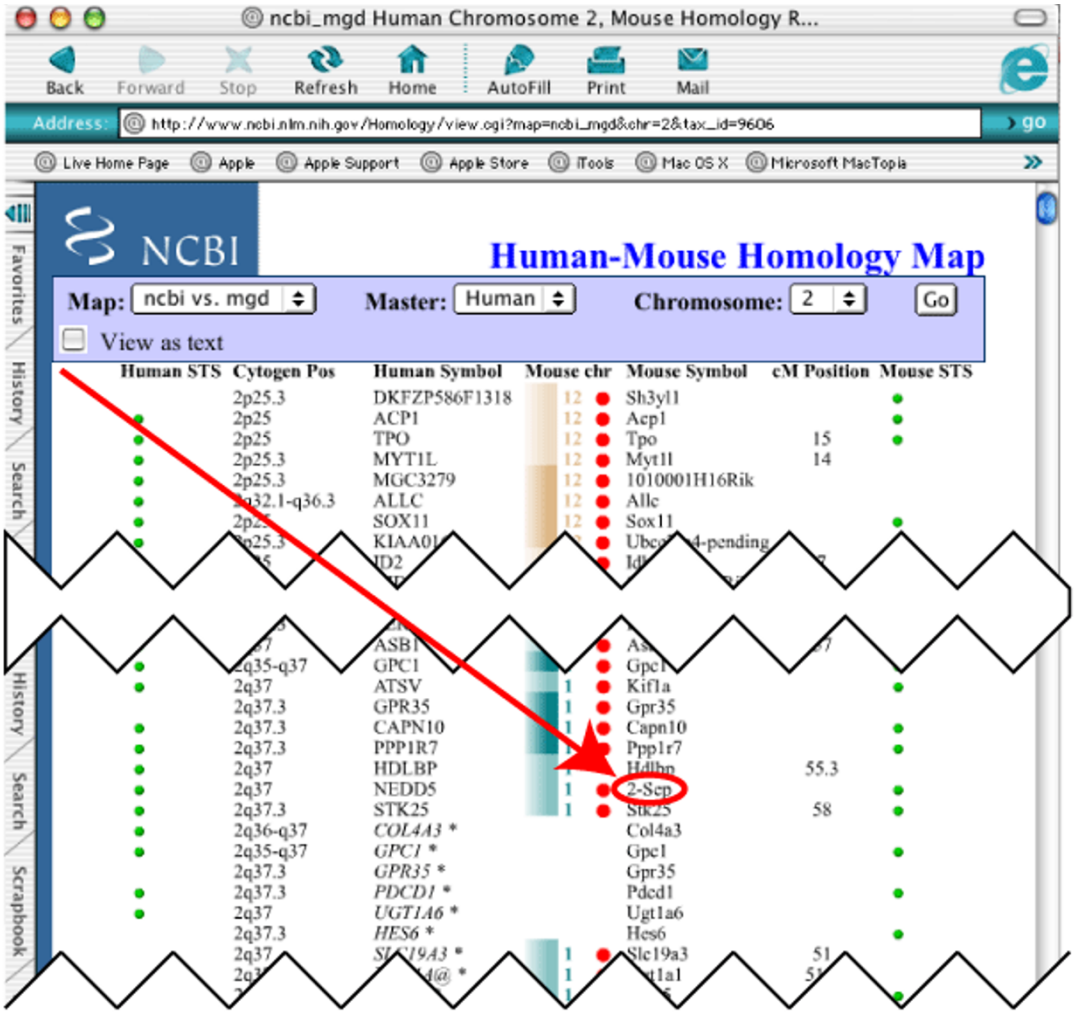


Figure 3
Screen shot of the human-mouse homology map from November 14, 2003 illustrating an error caused by default conversion of a gene name to date.

DEC1, a possible target for cancer therapy, was incorrectly rendered, and it could potentially be missed in downstream data analysis. The same type of error can infect, and propagate through, the major public data resources. For example, this type of error occurs several times in even the immaculately curated LocusLink database (Figure 2). This error in LocusLink originated from the human-mouse homology map data (Figure 3).

There are a number of work-arounds to these behaviors in Excel, but all of them require continued attention on the part of the user to avoid introducing errors. The appropriate solutions depend on the context in which Excel is opened:

If Excel is configured to open a test file automatically from another application, then the data must be pre-processed in the upstream application. For example, a space character or an apostrophe can be placed in front of the gene name. That is the solution implemented in the Excel output format option of MatchMiner [1] and the primary

```
#! /bin/csh

foreach arg ($*)
echo $arg

tr "\r" "\n" < $arg | sed '1,$s/\^M/\\
/g' | gawk
'(/[0-9]\-
(JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC|Jan|Feb|
Mar|Apr|M
ay|Jun|Jul|Aug|Sep|Oct|Nov|Dec|jan|feb|mar|
apr|may|jun|jul|aug|sep|oct|nov|dec)/ ||
apr|may|jun|jul|aug|sep|oct|nov|/[0-9]\.[0-9][0-9]E\+[[0-
9][0-9]/)
{print NR,$0}'
end
```

Figure 4
Script to scan for SymbolMutation error.

approach recommended by Microsoft in their Knowledge Base Article on the issue [6].

If a text file is to be opened by Excel, open Excel first and then select the text file to read. Then select "text" mode for the column(s) containing potentially affected symbols in the Text Import Wizard Step 3 of 3.

If text is to be copied from another application (such as a text processor) and pasted into a pre-opened Excel spreadsheet, the formatting must be set in the spreadsheet. Within the pre-opened spreadsheet, prior to pasting, use Format -> cells to specify which columns of the recipient spreadsheet are to be treated as text [6]. That procedure works for copying from several text processors tested on Mac OS 10.2. However, changing the format of the spreadsheet column to text fails to solve the conversion problem when pasting data from a Microsoft Word file. In that case, in addition to the formatting, use the Paste Special -> Paste: As: Text command to insert the text.

Despite the work-arounds, even the most vigilant investigator can inadvertently introduce conversion errors, and it is often necessary to screen data received from other sources. For that reason, we have provided the text of a Unix C shell program that detects possible gene-to-date and floating-point conversion problems (Figure 4). A downloadable version is available in the Supplementary Materials (see Additional file 1) and on our companion web site [7]. We have implemented a version in MatchMiner [1] and are releasing a version of GoMiner [2] that includes a quality check that uses the script.

We hope the date and floating-point conversions will be made non-default options – in deference to the large bioinformatics and biotechnology communities if not for other users. Even if that is done, however, there will be a lag time before all researchers have the new program version and an even longer time of confusion before the existing errors and inconsistencies have been expunged from all public and private databases. In the meantime, it is important to be alert to the problem.

The twin aims of this paper are (i) to minimize additional contamination of the public databases and literature, and (ii) to provide a cautionary note to the many researchers who put their microarray data through the exceptionally useful medium of an Excel spreadsheet, then map the clone identifiers or GenBank accession numbers into human-readable gene names. There is no way to know how many times and in how many laboratories the default date and floating point conversions to non-gene

names have adversely affected an experiment or caused genes to "disappear" from view.

Additional material

Additional File 1

Symbol Mutation Scan: Shell file in downloadable format Companion Website: http://discover.nci.nih.gov/symbolmutation
Click here for file [NOTE: to correct an error, this file was updated on 20 August 2004]
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-80-S1.sh]

Acknowledgements

We thank Dr. Keith Collins (Laboratory of Biosystems and Cancer, CCR, NCI, NIH) and Jocelyn Hsu (SRA International) for helpful comments and technical review of the manuscript.

References

1. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay W, Weinstein JN: **MatchMiner: a tool for batch navigation among gene and gene product identifiers.** *Genome Biol* 2003, **4:**R27.
2. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4:**R28.
3. Yun Z, Maecker HL, Johnson RS, Giaccia AJ: **Inhibition of PPAR gamma 2 gene expression by the HIF-1-regulated gene DEC1/Stra13: a mechanism for regulation of adipogenesis by hypoxia.** *Dev Cell* 2002, **2:**331-341.
4. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y, FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
5. Konno H, Fukunishi Y, Shibata K, Itoh M, Carninci P, Sugahara Y, Hayashizaki Y: **Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library.** *Genome Res* 2001, **11:**281-289.
6. **Microsoft Knowledge Base Article 21 XL: Text or Number Converted to Unintended Number Format** [http://support.microsoft.com/default.aspx?scid=kb;EN-US;Q214233]
7. **Mistaken Identifiers Companion** [http://discover.nci.nih.gov/symbolmutation]