Gene Updater: A Streamlit web tool that autocorrects and updates for Excel misidentified gene names

Clara WT Koh
Duke-NUS Medical School
Justin SG Ooi
Duke-NUS Medical School
Gabrielle LC Joly
Duke-NUS Medical School
Kuan Rong Chan (kuanrong.chan@duke-nus.edu.sg)
Duke-NUS Medical School

Research Article

Keywords: Transcriptomics, Gene expression studies, Streamlit, Web tool, Excel, Data Science, Python, Date-to-gene converter

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1146062/v1

Abstract

Background

Opening and processing gene expression data files in Excel runs into the inadvertent risk of converting gene names to dates. A plausible solution is to update these genes and dates to the new approved gene names as recommended by the HUGO Gene Nomenclature Committee (HGNC).

Results

We found that molecular pathways related to cell division, exocytosis, cilium assembly, protein ubiquitination and nitric oxide biosynthesis are most affected by this Excel auto-conversion. To circumvent this issue, we developed a web tool, Gene Updater, with Streamlit that can convert old gene names and dates back into the new gene names recommended by HGNC. The running instance of the web tool is accessible at: https://share.streamlit.io/kuanrongchan/date-to-gene-converter/main/date_gene_tool.py

Conclusions

Gene Updater can convert old gene names and dates back into the updated gene names, which are more resilient to Excel auto-conversion. We envision this tool to facilitate the sharing of gene expression datasets across multiple analytics platforms.

Background

When gene expression datasets are opened with Excel under default settings (Microsoft Corp., Redmond, WA), a recurring problem where gene names are converted to dates occurs. Similarly, if gene names are copied from another application (e.g. text processors) and pasted into an Excel spreadsheet without specifying cell formatting, conversion of gene names to dates can occur [1]. This can cause voids in pathway enrichment analyses, as the genes that are converted dates are not recognised in pathway databases. For instance, septins (e.g. SEPT1) which are involved in cell division are internally converted to SEP-01in Excel, which cannot be recognised by pathway databases such as Enrichr and Gene-set enrichment analysis (GSEA). This problem has become so rampant that approximately one-fifth of the published papers with supplementary Excel gene lists contain erroneous gene name conversions [2, 3]. As many of these datasets are frequently accessed by other data scientists, such errors may be carried over to other scientific publications, resulting in further distortion of downstream data analysis.

To tackle this issue, the HUGO Gene Nomenclature Committee (HGNC) announced in 2017 to update the gene names that may be unintentionally converted to dates in Excel files [4]. This movement was well-received by researchers and data scientists, as changing to the updated gene names would allow sharing

of gene expression data without worrying about the automatic conversion of gene symbols to dates in Excel. However, at present, the majority of the published gene expression data are not updated to the new approved gene names, especially in microarray datasets. We thus developed a Gene Updater web tool that allows researchers to convert the previous gene names to the new approved gene names recommended by HGNC. Moreover, in the event where the gene names are unintentionally converted to dates by Excel, the web tool allows researchers to convert these terms back to the updated approved gene names. We believe that these efforts will facilitate gene expression data sharing between researchers who may be working on different analytics platforms.

Implementation

Development of Gene Updater

We used Streamlit (https://www.streamlit.io), which is an open-source app framework using the Python programming language. The GitHub codes that are used to build the Gene Updater can be accessed at https://github.com/kuanrongchan/date-to-gene-converter. As we noticed that there is also a possibility that Excel may convert genes to dates in other formats, such as dd/mm/yyyy, yyyy/mm/dd formats, additional instructions are provided to the users for conversion of these date formats to the new updated genes. With the codes available in GitHub, users can choose to run the Gene Updater locally. However, to facilitate the non-bioinformaticians, we have deployed the app using Streamlit Teams and the app is hosted at: https://share.streamlit.io/kuanrongchan/date-to-gene-converter/main/date_gene_tool.py

Results

Voids in pathway analyses when genes are converted to dates in Excel

The converted human gene names in Excel, with their respective floating-point numbers and updated approved gene names are displayed in Table 1. Pathway enrichment analysis [5] revealed that these genes play a critical role in cell division, exocytosis, cilium assembly, ubiquination and nitric oxide biosynthesis (Figure 1a). The genes involved in the respective biological pathways are as displayed in Figure 1a. Overall, these results highlight the potential voids in pathway enrichment analyses if the gene names are auto-converted to dates in Excel.

Table 1

Human gene names that are most frequently converted to dates in Excel. The respective updated gene name and gene description is also provided.

Previous Gene Name	Excel Date Conversion	HUGO Gene	Entrez Gene Description (Homo sapiens)
DEC1	Dec-01	DELEC1	deleted in esophageal cancer 1
MARC1	Mar-01	MTARC1	mitochondrial amidoxime reducing component 1
MARCH1	Mar-01	MARCHF1	membrane associated ring finger 1
MARCH2	Mar-02	MARCHF2	membrane associated ring finger 2
MARC2	Mar-02	MTARC2	mitochondrial amidoxime reducing component 2
MARCH3	Mar-03	MARCHF3	membrane associated ring finger 3
MARCH4	Mar-04	MARCHF4	membrane associated ring finger 4
MARCH5	Mar-05	MARCHF5	membrane associated ring finger 5
MARCH6	Mar-06	MARCHF6	membrane associated ring finger 6
MARCH7	Mar-07	MARCHF7	membrane associated ring finger 7
MARCH8	Mar-08	MARCHF8	membrane associated ring finger 8
MARCH9	Mar-09	MARCHF9	membrane associated ring finger 9
MARCH10	Mar-10	MARCHF10	membrane associated ring finger 10
MARCH11	Mar-11	MARCHF11	membrane associated ring finger 11
SEPT1	Sep-01	SEPTIN1	septin 1
SEPT2	Sep-02	SEPTIN2	septin 2
SEPT3	Sep-03	SEPTIN3	septin 3
SEPT4	Sep-04	SEPTIN4	septin 4
SEPT5	Sep-05	SEPTIN5	septin 5
SEPT6	Sep-06	SEPTIN6	septin 6
SEPT7	Sep-07	SEPTIN7	septin 7
SEPT8	Sep-08	SEPTIN8	septin 8
SEPT9	Sep-09	SEPTIN9	septin 9
SEPT10	Sep-10	SEPTIN10	septin 10
SEPT11	Sep-11	SEPTIN11	septin 11

Previous Gene Name	Excel Date Conversion	HUGO Gene	Entrez Gene Description (Homo sapiens)
SEPT12	Sep-12	SEPTIN12	septin 12
SEPT14	Sep-14	SEPTIN14	septin 14
SEP15	Sep-15	SELENOF	15 kDa selenoprotein

Developing Gene Updater to convert dates and old gene names to the updated gene names

We next explored the use of Streamlit (https://www.streamlit.io) to build an interactive web tool, termed as Gene Updater, that converts floating-point numbers and the old gene names to the new gene names. The running instance of the online app is deployed at https://share.streamlit.io/kuanrongchan/date-to-gene-converter/main/date_gene_tool.py and the GitHub codes to build the web tool can be accessed at https://github.com/kuanrongchan/date-to-gene-converter.

The user interface starts with a file uploader that enables users to upload their .csv or .xlsx file(s). Multiple files can be uploaded and the first column should contain the gene names for gene conversion to occur. The checkbox at the side-bar allows users to inspect the dataframe that is successfully uploaded. In addition, to allow users to explore the functionalities of the tool, a demo dataset displaying an Excel file that contains dates on the first column and the corresponding gene description on the second column is pre-loaded.

If old gene names are uploaded onto the web tool, then all the gene names will be changed to the updated one. If dates are provided on the first column, all dates can be converted to the updated gene names with the web tool, with the exception of Mar-01 and Mar-02 as these dates can be potentially mapped to more than one gene (Figure 1c).

Converting dates which can be mapped to more than one gene

As shown in Table 1, Mar-01 can be mapped to either MTARC1 and MARCHF1. Similarly, Mar-02 can be mapped to either MTARC2 and MARCHF2. It is critical not to map these genes wrongly as the functions of these genes are disparate. The Gene Updater web tool is hence designed to allow users to assign the correct gene names to the respective date terms. If the data file contains the gene description column, then the gene names can be most easily assigned by the user with the web tool. However, if gene description is not included in the dataset, we have left the responsibility of the user to check with the raw data file to correctly assign the duplicate terms. After manually assigning MTARC1, MARCHF1, MTARC2 and MARCHF2 to the respective rows, the converted dataframe with the updated gene list can be downloaded and exported for other downstream data analysis. Users can also use the multiquery search

bar within the web tool to inspect that the gene names are successfully updated before downloading the converted dataset.

Discussion

The automatic conversion of gene names to dates poses a problematic feature of Excel. Analysis of major journals revealed that over 50% of journals with supplementary Excel files had at least one file with such errors [2]. At present, there is no way to permanently deactivate the autoconversion. The fastest way to spot these errors is by sorting the column of gene names in ascending order, and the gene symbols that are converted to dates will appear as numbers at the top of the column. Thereafter, users will often have to manually convert these dates to text by specifying the cell formatting within Excel, which can be tedious and potentially error-prone. This date-to-gene converter web tool hence allows researchers to convert either old gene names or dates to the updated gene names quickly and reproducibly. In addition, we have incorporated multiple checkpoints that will allow users to inspect the data before and after conversion.

Streamlit is an open-source framework that allows developers to create and deploy web apps easily. As the codes are also publicly available at GitHub, developers can easily customise the codes to convert and update any terms of interest. At present, the two tools that can potentially convert dates to gene names are Oct4th (https://oct4th.sandbox.bio) and Truke (http://maplab.imppc.org/truke/) [6]. However, Oct4th can only work on gene data files that have not been manipulated and processed in Excel. Moreover, the tool is presently unable to convert to the updated gene names, which are more resilient to auto-conversion. Truke can potentially convert the date formats to gene names, but can only convert dates that are labelled in the dd/mm/yy format, and process files one at a time. In contrast, our Gene Updater Streamlit web tool can process multiple .csv and .xlsx files, and takes into account the different kinds of date formatting that are converted by Excel, thus allowing faster and more efficient processing of dates to genes.

Conclusion

In summary, we developed a publicly available, user-friendly and customisable web tool that converts old gene names and dates back into updated gene names. We strongly encourage processing of gene expression datasets with this web tool before publication or sharing of genetic information, to minimise the risk of gene to date auto-conversion.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Project name: Gene Updater

Project homepage: https://github.com/kuanrongchan/date-to-gene-converter

Operating systems: Platform independent.

Programming language: Python

Other requirements: None. If web tool is run locally, Python3 is recommended

Licence: None

Restrictions to non-academics: None

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Individual Research Grant (MOH-000610).

Authors' contributions

C.W.T.K., J.S.G.O., K.R.C., developed the web tool. G.C.W.L. and C.W.T.K. did the pathway analysis. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

We would like to thank Eugenia Ong Ziying and Darren Mok Zhi Liang for the genuine feedback on the web tool.

Abbreviations

HGNC: HUGO Gene Nomenclature Committee

MARC1/MTARC1: Mitochondrial amidoxime reducing component 1

MARCH1/MARCHF1: Membrane associated ring finger 1

MARC2/MTARC2: Mitochondrial amidoxime reducing component 2

MARCH2/MARCHF2: Membrane associated ring finger 2

References

1. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, Barrett JC, Weinstein JN: **Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics**. *BMC Bioinformatics* 2004, **5**:80.
2. Abeysooriya M, Soria M, Kasu MS, Ziemann M: **Gene name errors: Lessons not learned**. *PLoS Comput Biol* 2021, **17**:e1008984.
3. Ziemann M, Eren Y, El-Osta A: **Gene name errors are widespread in the scientific literature**. *Genome Biol* 2016, **17**:177.
4. Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S: **Guidelines for human gene nomenclature**. *Nat Genet* 2020, **52**:754–758.
5. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool**. *BMC Bioinformatics* 2013, **14**:128.
6. Mallona I, Peinado MA: **Truke, a web tool to check for and handle excel misidentified gene symbols**. *BMC Genomics* 2017, **18**:242.

Figures

Figure 1

a. Top 10 enriched pathways based on genes that are frequently converted to dates in Excel. Genes were analysed against the GO Biological Processes database with Enrichr, and all presented pathways had adjusted p-values < 0.05. b. Schematic of the web tool function. If previous gene names are provided, these genes will be automatically converted to the updated approved gene names. If dates are provided, all genes, with the exception of MAR-01 and MAR-02 will be converted to the new approved gene names. For MAR-01 and MAR-02, users will need to do an additional step to assign the genes to either MTARC1, MARCHF1, MTARC2 or MARCHF2.