

Keemei: cloud-based validation of tabular bioinformatics file formats in Google Sheets

Jai Ram Rideout¹, John H. Chase¹, Evan Bolyen¹, Gail Ackermann², Antonio González², Rob Knight^{2,3}, J. Gregory Caporaso^{1,4,*}

¹ Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, Arizona, USA.

² Department of Pediatrics, University of California San Diego, San Diego, California, USA.

³ Department of Computer Science & Engineering, University of California San Diego, San Diego, California, USA.

⁴ Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona, USA.

* Corresponding author.

Author e-mail addresses:

Jai Ram Rideout: jai.rideout@gmail.com

John H. Chase: chase.johnh@gmail.com

Evan Bolyen: ebolyen@gmail.com

Gail Ackermann: ackermag@ucsd.edu

Antonio González: antgonza@gmail.com

Rob Knight: robknight@ucsd.edu

J. Gregory Caporaso: gregcaporaso@gmail.com

Abstract

Background

Bioinformatics software often requires human-generated tabular text files as input and have specific requirements for how those data are formatted. Users frequently manage these data in spreadsheet programs, which is convenient for researchers who are compiling the requisite information because the spreadsheet programs can easily be used on different platforms including laptops and tablets, and because they provide a familiar interface. It is increasingly common for many different researchers to be involved in compiling these data, including study coordinators, clinicians, lab technicians, and bioinformaticians. As a result, many research groups are shifting toward using cloud-based spreadsheet programs, such as Google Sheets, which support concurrent editing of a single spreadsheet by different users working on different platforms. Often most of the researchers who are entering data will not be familiar with the formatting requirements of the bioinformatics programs that will be used, so validating and correcting file formats is often a bottleneck prior to beginning bioinformatics analysis.

Findings

We present Keemei, a Google Sheets Add-on for validating tabular files used in bioinformatics analyses. Keemei is available free of charge from Google's Chrome Web Store. Keemei can be installed and run on any web browser supported by Google Sheets. Keemei currently supports validation of two widely used tabular bioinformatics formats, the QIIME sample metadata

mapping file format, and the Spatially Referenced Genetic Data (SRGD) format, but is designed to easily support the addition of others.

Conclusions

Keemei will save researchers time and frustration by providing a convenient interface for tabular bioinformatics file format validation. By allowing everyone involved with data entry for a project to easily validate their data, it will reduce the validation and formatting bottlenecks that are commonly encountered when human-generated data files are first used with a bioinformatics system. Simplifying the validation of essential tabular data files, such as sample metadata, will reduce common errors and thereby improve the quality and reliability of research outcomes.

Keywords

data validation, tabular file format, spreadsheet, QIIME, metadata, cloud

Findings

Many bioinformatics applications require human-generated tabular text files as input and have specific requirements for how those data are formatted. A common example is metadata describing a collection of biological samples (i.e., a *sample-metadata mapping file*), such as ISA-TAB-based file formats [1]. For example, in a human microbiome survey, this file would map unique sample identifiers [2] to descriptions of each sample for minimum information standards compliance [3], study-specific parameters such as host identifier and disease state, and technical information such as, for marker genes, the PCR primer pair that was used to amplify and sequence the gene reading out the community profile, and, for shotgun metagenomics, the

library construction protocol. These data are generally compiled by different people who typically differ in their knowledge of the requirements of the bioinformatics analysis tools or who may not even know which bioinformatics tools will be used, and who lack complete information about the end-to-end study design. For example, a study coordinator may compile per-subject demographic information, a clinician may compile medical information, a lab technician may compile information about the DNA extraction and sequencing, and a bioinformatician may compile any missing minimum-standards-compliance information. As a result, the most time-consuming step in a bioinformatics analysis is often merging these data from different human-generated sources and bringing them into compliance with the format specifications of the bioinformatics program(s) that will be used.

Users generally manage their tabular data in spreadsheet programs (e.g., Microsoft Excel). This is convenient for researchers who are compiling the requisite information (the study coordinators, clinicians, etc.) because the spreadsheet programs can easily be used on different platforms including laptops and tablets, and because they provide a familiar interface. A common issue arises, however, when multiple people are responsible for compiling different information for a tabular document. Versions of the document rapidly become out-of-sync, for example if a clinician and a study coordinator are adding information at the same time, or if one person accidentally adds information to an outdated version of the file. Cloud-based spreadsheet programs that allow concurrent editing, such as Google Sheets [4], can alleviate these issues because there is always one definitive version of the document that can be edited by multiple users at the same time. For this reason, among others, Google Sheets is becoming increasingly popular for creating, editing and managing human-generated tabular files used in bioinformatics analyses.

Often a bioinformatics package will include a file format validator as part of its suite of tools, but validating files is often cumbersome. The user will typically export their tabular data from their spreadsheet program in the format expected by the validator (e.g., CSV or TSV), run the validator, and then return to their spreadsheet program to correct errors. After correcting errors, they would again export their data from the spreadsheet program, re-run the validator, and repeat the process until no more errors are present. If data needed to be corrected or added to the tabular data file at some point in the future, the process would be repeated (and all old versions of that file would need to be updated). For example, this is how the sample metadata mapping file validation workflow has traditionally been performed for QIIME through versions 1.9.1 [5] (QIIME is a widely used bioinformatics package for microbiome analysis that is developed and maintained by the authors of this paper, among others). In addition to being slow, this workflow can easily result in many different versions of the sample metadata mapping file, which frequently leads to confusion about which is the latest or definitive version of the file.

We present Keemei, a Google Sheets Add-on for validating tabular files used in bioinformatics analyses. Keemei is available free of charge from Google's Chrome Web Store [6]. Keemei can be installed and run on any web browser supported by Google Sheets, so browser support is not limited to Google Chrome. After installation, users are provided with a new menu option in their Google Sheets to perform tabular data validation for specific bioinformatics file formats. When a user validates their tabular data, a report in a sidebar indicates whether there are any errors or warnings in the file, and if so, which cells contain errors or warnings. Invalid cells are also highlighted directly in the spreadsheet, and hovering the mouse over a cell will display the reason(s) why the cell is invalid (Figure 1). The user can click on a cell to view the reason(s)

why the cell is invalid, and also has the option to navigate directly to a cell in the spreadsheet in order to fix it. This feature is especially useful for navigating large spreadsheets that would require scrolling to find and correct invalid cells (Figure 2). The user can correct invalid cells and then re-validate the file, repeating the process as necessary until there are no more errors or warnings. While this process is iterative, it all occurs within the Google Sheets interface, avoiding repetitive, time-consuming, and error-prone importing and exporting of files for external validation. Since there is no need to export to a new file each time the validator is run, a single definitive spreadsheet can be maintained throughout the lifetime of a project, avoiding multiple versions of exported spreadsheets.

Building Keemei as a Google Sheets Add-on provides several additional benefits over a stand-alone validator, or a plugin for a non-cloud-based spreadsheet program such as Microsoft Excel. First, as noted above, cloud-based spreadsheet programs that allow concurrent editing by multiple users assist with keeping versions of files synchronized. Building Keemei on top of a cloud-based program therefore allows for validation of tabular file formats in the same interface that we recommend to be used for data entry and correction of errors. Next, Keemei is largely platform independent: it can be used on any system that can run Google Sheets, and doesn't, for example, require installing the bioinformatics software that will ultimately be used for data analysis. This cloud-based mechanism of interacting with software is increasingly popular in bioinformatics, as installation is typically trivial or not required, the burden of maintenance and upgrades is shifted from the user to the developer, and in many cases it results in a graphical interface for software that previously had only a command line interface. This is exemplified in the many applications that now support Galaxy wrappers [7]. Next, Google Sheets has built-in versioning support so that it is possible to revert to previous versions of the spreadsheet. This is

useful for determining if or when errors may have been introduced into tabular data. Users also won't need to install new versions of Keemei as it is released. When the developers push a new version to the Chrome Web Store, it is automatically updated in users' Google Sheets environments. Finally, there are many other relevant tools being developed and released for Google Sheets, so Keemei users will have access to other useful functionality within this interface. For example, users can easily obtain graphical summaries of their data using the *Explore* function that is built into spreadsheets (e.g., a patient age histogram can automatically be generated from their sample metadata mapping files), or tag their metadata with relevant ontology terms using OntoMaton [8].

There are a couple of drawbacks to developing Keemei as a Google Sheets Add-on that should be noted. First, Google limits the size of spreadsheet that can be loaded (at the time of writing, around 2 million cells). Depending on the data to be validated, this may or may not be a problem. Next, being cloud-based, it's possible that Institutional Review Boards (IRBs) or other ethics or data management committees may disallow the use of Keemei for studies involving human subjects research or confidential research, even if the data is not made public.

Researchers should discuss the use of Google Sheets with their IRB or other relevant bodies prior to starting their study. One step that could be taken to alleviate IRB concerns would be to ensure that no personally identifying information is contained in the data files loaded in Google Sheets.

Keemei currently supports validation of two specific tabular bioinformatics file formats: the QIIME sample metadata mapping file [9], and the Spatially Referenced Genetic Data (SRGD, also known as SRGD.csv) file. Both of these are used and/or generated by multiple

bioinformatics programs, including QIIME, geneGIS [10], and Wildbook [11]. Keemei is designed to support the inclusion of additional format validators, and others will be added in the future.

Keemei will save researchers time and frustration by providing a convenient interface for tabular bioinformatics file format validation in a spreadsheet interface they are already familiar with. It will easily allow everyone involved with data entry for a project to perform validation of their data, reducing the validation and formatting bottleneck that is often encountered when human-generated data files are first used with a bioinformatics system. We additionally hope that the availability of Keemei and other Google Sheets Add-ons such as OntoMaton will encourage a shift away from locally installed software for data management and processing toward cloud-based solutions where multiple users can access and operate on the same files at the same time. Simplifying the tracking of essential tabular data files, such as sample metadata, will reduce common errors and thereby improve the quality and reliability of research outcomes.

Availability and requirements

Project name: Keemei

Project home page: <http://keemei.qiime.org>

Operating system(s): Platform independent

Programming language: Google Apps Script, JavaScript/HTML/CSS

Other requirements: Web browser supported by Google Sheets

License: BSD 3-Clause

Any restrictions to use by non-academics: None

List of abbreviations used (if any)

PCR: Polymerase Chain Reaction

QIIME: Quantitative Insights Into Microbial Ecology

SRGD: Spatially Referenced Genetic Data

Competing interests

The authors declare no competing interests.

Authors' contributions

JRR and EB designed and developed Keemei. JC, EB, GA, AGP, RK and JGC tested the system and provided feedback on features and functionality. JC designed the Keemei logo. JRR and JGC wrote the manuscript.

Acknowledgements

The authors wish to thank Yoshiki Vazquez-Baeza and Adam Robbins-Pianka for helpful suggestions and discussion during the development of Keemei.

References

1. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, Neumann S, Sterk P, Tong W, Sansone S-A: **ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community**

level. *Bioinformatics* 2010, **26**:2354–2356.

2. Chase JH, Bolyen E, Rideout JR, Gregory Caporaso J: **cual-id: Globally Unique, Correctable, and Human-Friendly Sample Identifiers for Comparative Omics Studies.** .

3. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, et al.: **Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications.** *Nat Biotechnol* 2011, **29**:415–420.

4. **Google Sheets** [<http://www.google.com/sheets>]

5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunencko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods* 2010, **7**:335–336.

6. **Chrome Web Store** [<https://chrome.google.com/webstore/category/apps>]

7. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86.

8. Maguire E, González-Beltrán A, Whetzel PL, Sansone S-A, Rocca-Serra P: **OntoMaton: a bioportal powered ontology widget for Google Spreadsheets.** *Bioinformatics* 2013, **29**:525–527.

9. **QIIME File Format Descriptions** [http://qiime.org/documentation/file_formats.html]

10. Dick DM, Walbridge S, Wright DJ, Calambokidis J, Falcone EA, Steel D, Follett T, Holmberg J, Baker CS: **geneGIS: Geoanalytical Tools and Arc Marine Customization for Individual-Based Genetic Records.** *Trans GIS* 2014, **18**:324–350.

11. **Wildbook Framework for Mark-Recapture Studies**
[<http://www.wildme.org/wildbook/doku.php>]

12. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R: **Moving pictures of the human microbiome.** *Genome Biol* 2011, **12**:R50.

13. Lauber CL, Hamady M, Knight R, Fierer N: **Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale.** *Appl Environ Microbiol* 2009, **75**:5111–5120.

Illustrations and figures

Figure 1. Keemei screenshots of a user validating a QIIME sample metadata mapping file. From the Google Sheets menu bar, the user selects `Add-ons > Keemei > Validate QIIME mapping file` to validate their spreadsheet against the QIIME sample metadata mapping file format specification. Validation results are displayed directly in the spreadsheet and in a sidebar interface on the right side of the spreadsheet. Invalid cells are highlighted in the spreadsheet, where a red background color indicates the cell has one or more errors or warnings associated with it, and a yellow background color indicates the cell has one or more warnings associated with it. Hovering the mouse over a cell will display the reason(s) why the cell is invalid. The sidebar contains a summary of the validation (e.g., file format validated against, number of invalid cells, etc.) and lists invalid cells. The user can click on a cell in the sidebar to view the reason(s) why the cell is invalid (similar to hovering over a cell in the spreadsheet). In this screenshot, the user has clicked on invalid cell A3; we see that cells A3 and A5 contain duplicate sample identifiers, which are disallowed in QIIME mapping files. The data in this screenshot are derived from [12].

Keemei demo ☆

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive Comments Share

Keemei

- Validate QIIME mapping file
- Validate SRGD file
- Clear validation status
- About
- Help

#SampleID	A	B	Barcode	SampleType	Year	Count
1	#SampleID	LinkerPrimerSequence	Barcode			
2	L1S8	GTGCCAGCMGCCGCGGTAA	AGCTGACTAGTC	gut		
3	L1S140	GTGCCAGCMGCCGCGGTAA	ATGGCAGCTCTA	gut		
4	L1S57	GTGCCAGCMGCCGCGGTAA	ACACACTATGGC			
5	L1S140	GTGCCAGCMGCCGCGGTAA	CTGAGATACGCG	gut	2009	1
6	L1S76	GTGCCAGCMGCCGCGGTAA	ACTACGTGTGGT	gut	2009	2
7	L1S-105	GTGCCAGCMGCCGCGGTAA	AGCTGACTAGTC	gut	2009	3
8	L1S257	GTGCCAGCMGCCGCGGTAA	CCGACTGAGATG	gut	2009	3
9	L1S281	GTGCCAGCMGCCGCGGTAAZ	CCTCTCGTGATC	gut	2009	4
10	L2S240	GTGCCAGCMGCCGCGGTAA	CATATCGCAGTT	left palm	2008	10
11	L2S155	GTGCCAGCMGCCGCGGTAA	ACGATGCGACCA	left palm	2009	1
12	L2S309	GTGCCAGCMGCCGCGGTAA	CGTGCATTATCA	left palm	2009	1
13	L2S175	GTGCCAGCMGCCGCGGTAA	AGCTATCCACGA	left palm	2009	2
14	L2S204	GTGCCAGCMGCCGCGGTAA	ATGCAGCTCAGT	left palm	2009	3
15	L2S357	GTGCCAGCMGCCGCGGTAA	CTAACGCAGTCAT	left palm	2009	3
16	L2S222	GTGCCAGCMGCCGCGGTAA	CACGTGACATGT	left palm	2009	4
17	L2S382	GTGCCAGCMGCCGCGGTAA	CTCAATGACTCA	left palm	2009	4

QIIME Illumina Overview Tutorial QIIME 88 Soils S Explore



Keemei demo ☆

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive Comments Share

Keemei validation report

17 invalid cells
12 errors
6 warnings

Invalid cells

- A1
- A3
- A5
- A7
- B1

Errors:

- Duplicate sample ID. Duplicates in A3, A5

#SampleID	A	B	C	D
1	#SampleID	LinkerPrimerSequence	BarcodeSequence	SampleType
2	L1S8	GTGCCAGCMGCCGCGGTAA	AGCTGACTAGTC	gut
3	L1S140	ERRORS:	GCAGCTCTA	gut
4	L1S57	Duplicate sample ID. Duplicates in A3, A5	CACTATGGC	
5	L1S140		AGATACGCG	gut
6	L1S76		ACGTGTGGT	gut
7	L1S-105	GTGCCAGCMGCCGCGGTAA	AGCTGACTAGTC	gut
8	L1S257	GTGCCAGCMGCCGCGGTAA	CCGACTGAGATG	gut
9	L1S281	GTGCCAGCMGCCGCGGTAAZ	CCTCTCGTGATC	gut
10	L2S240	GTGCCAGCMGCCGCGGTAA	CATATCGCAGTT	left palm
11	L2S155	GTGCCAGCMGCCGCGGTAA	ACGATGCGACCA	left palm
12	L2S309	GTGCCAGCMGCCGCGGTAA	CGTGCATTATCA	left palm
13	L2S175	GTGCCAGCMGCCGCGGTAA	AGCTATCCACGA	left palm
14	L2S204	GTGCCAGCMGCCGCGGTAA	ATGCAGCTCAGT	left palm
15	L2S357	GTGCCAGCMGCCGCGGTAA	CTAACGCAGTCAT	left palm
16	L2S222	GTGCCAGCMGCCGCGGTAA	CACGTGACATGT	left palm
17	L2S382	GTGCCAGCMGCCGCGGTAA	CTCAATGACTCA	left palm

QIIME Illumina Overview Tutorial QIIME 88 Soils S Explore

Figure 2. **Keemei screenshots of a user focusing on an invalid cell in a QIIME mapping file.** Keemei's sidebar provides a way to focus on an invalid cell in order to correct it. This feature is especially useful when working with large sheets that would require scrolling to find and correct invalid cells. By clicking on the magnifying glass next to invalid cell O46 in the sidebar (red border added for clarity), cell O46 is made active and the user is scrolled to the cell's location in the spreadsheet to make any necessary corrections. The data in this screenshot are derived from [13].

Keemei demo ☆

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive Comments Share

Keemei validation report

Validation report for sheet **QIIME 88**


Soils


QIIME mapping file format

Summary

2 invalid cells
0 errors
2 warnings

Invalid cells

▶ O46 

▶ O66 

Close

#SampleID	A	B	C	D	E
1	#SampleID	BarcodeSequence	LinkerPrimerSeq	TARGET_SUBF	ASSIGNED_FF
2	IT2.141720	ACGTGCCGTAC	CATGCTGCCTC	V2	n
3	HI3.141676	ACGCTATCTGC	CATGCTGCCTC	V2	n
4	MD2.141689	ACTCGATTCPA	CATGCTGCCTC	V2	n
5	CA1.141704	ACACGAGCCAC	CATGCTGCCTC	V2	n
6	PE5.141692	AGACTGCGTAC	CATGCTGCCTC	V2	n
7	CO1.141714	ACATGATCGTT	CATGCTGCCTC	V2	n
8	DF3.141696	ACCGCAGAGTC	CATGCTGCCTC	V2	n
9	PE1.141715	ACTTGTAGCAC	CATGCTGCCTC	V2	n
10	SP2.141678	AGCGCTGATG	CATGCTGCCTC	V2	n
11	CO3.141651	ACATTCAGCGC	CATGCTGCCTC	V2	n
12	SA2.141687	AGATCGGCTCC	CATGCTGCCTC	V2	n
13	CM1.141723	ACATCACTTAG	CATGCTGCCTC	V2	n
14	LQ2.141729	ACTCACGGTAT	CATGCTGCCTC	V2	n
15	SR2.141673	AGCTATCCACC	CATGCTGCCTC	V2	n
16	CR1.141682	ACCACATACAT	CATGCTGCCTC	V2	n
17	VC1.141722	AGTGTGATCC	CATGCTGCCTC	V2	n

QIIME Illumina Overview Tutorial QIIME 88 Soils S Explore



Keemei demo ☆

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive Comments Share

Keemei validation report

Validation report for sheet **QIIME 88**


Soils


QIIME mapping file format

Summary

2 invalid cells
0 errors
2 warnings

Invalid cells

▶ O46 

▶ O66 

Close

	L	M	N	O	P
31	14.2	soil metagenome		60 Santa Fe, NM U	1500
32	59.5	soil metagenome		32 Manu National P	2000
33	31	soil metagenome		65 Badlands Nation	1100
34	26.5	soil metagenome		54 Santa Barbara, C	5000
35	4.2	soil metagenome		58 Santa Fe, NM U	1500
36	42.5	soil metagenome		20 Sierra Nevada M	3000
37	22.9	soil metagenome		9 Sunset Crater, A	1900
38	3	soil metagenome		22 Sevilleta LTER, I	1480
39	27	soil metagenome		49 Institute for Ecos	7000
40	23.3	soil metagenome		35 Calhoun Experim	1500
41	69.5	soil metagenome		41 H.J. Andrews Ex	7000
42	68.9	soil metagenome		76 Konza Prairie LT	1000
43	5.7	soil metagenome		20 Mojave Desert, C	7700
44	45.9	soil metagenome		55 Sedgwick Reser	3000
45	158.8	soil metagenome		36 Kohala Peninsul	1000
46	107	soil metagenome		41 Mary's Peak, OR	1300
47	52.9	soil metagenome		52 Tonlik Lake LTER	8000

QIIME Illumina Overview Tutorial QIIME 88 Soils S Explore