DATABASE
The Journal of Biological Databases and Curation

Database, 2018, 1–8
doi: 10.1093/database/bay019
Original article

Original article

GEOMetaCuration: a web-based application for accurate manual curation of Gene Expression Omnibus metadata

Zhao Li[1,2], **Jin Li**[1,2] **and Peng Yu**[1,2,*]

[1]Department of Electrical and Computer Engineering and [2]TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA

*Corresponding author: Tel: 1-979-845-7441; Fax: 1-979-845-6259; Email: pengyu.bio@gmail.com

Citation details: Li,Z., Li,J. and Yu,P. GEOMetaCuration: a web-based application for accurate manual curation of Gene Expression Omnibus metadata. *Database* (2018) Vol. 2018: article ID bay019; doi:10.1093/database/bay019

Received 31 October 2017; Revised 30 January 2018; Accepted 30 January 2018

Abstract

Metadata curation has become increasingly important for biological discovery and biomedical research because a large amount of heterogeneous biological data is currently freely available. To facilitate efficient metadata curation, we developed an easy-to-use web-based curation application, GEOMetaCuration, for curating the metadata of Gene Expression Omnibus datasets. It can eliminate mechanical operations that consume precious curation time and can help coordinate curation efforts among multiple curators. It improves the curation process by introducing various features that are critical to metadata curation, such as a back-end curation management system and a curator-friendly front-end. The application is based on a commonly used web development framework of Python/Django and is open-sourced under the GNU General Public License V3. GEOMetaCuration is expected to benefit the biocuration community and to contribute to computational generation of biological insights using large-scale biological data. An example use case can be found at the demo website: http://geometacuration.yubiolab.org.

Database URL: https://bitbucket.com/yubiolab/GEOMetaCuration

Introduction

Metadata curation is an essential step for analyzing and integrating heterogeneous large-scale biological datasets generated by numerous labs across the world (1). Such data integration presents biomedical researchers with unprecedented opportunities to discover new biological insights that are hidden if each dataset is dealt with separately (2). However, it becomes increasingly challenging to curate semi-structured metadata efficiently and accurately, as the volume of biological data grows rapidly (3). For example, there are more than 100 000 datasets in one of the most popular public databases for functional genomics data, Gene Expression Omnibus (GEO) (4). It is critical to develop new methods to facilitate curation to ultimately enable streamlined biological discovery from a large amount of biological datasets.

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024

Despite the importance of metadata curation in biological discovery, the topic has received relatively little attention compared with other areas of data curation. Only recently, using techniques from natural language processing (NLP), Bernstein *et al.* (5) developed a computational pipeline to automatically recognize biomedical entities of human biological samples from metadata in the National Center for Biotechnology Information's (NCBI's) Sequence Read Archive (SRA). However, curating high-level summary information is still an unsolved problem in NLP (6), such as determining the key factors being perturbed in a dataset. Manually curating these metadata is essential. Indeed, a number of metadata resources, such as RNASeqMetaDB (7) and SFMetaDB (8), have been created using manual curation with high accuracy.

However, manual metadata curation is time-consuming. General-purpose applications, such as Google Drive, have been used to facilitate the curation process to a certain extent (7, 8). But the curation process still involves many mechanical and tedious operations interspersed throughout. Curators are required to collect information scattered in multiple places for each curation task to reach a correct conclusion. Much time is spent searching in these places to locate relevant fragments of text. In addition, due to potential errors made by each curator, a curation task must be assigned to multiple curators, and their curation results must be compared and integrated to ensure high curation accuracy. Such a coordinated effort can take a significant amount of time without the assistance of a tool.

To address these problems with metadata curation, we developed a curator-friendly web-based application, GEOMetaCuration, for GEO datasets. It was designed to eliminate mechanical steps in the curation process to improve curation productivity for a large amount of metadata. The application facilitates curation in two ways: 1) a well-designed back-end management system simplifies the process of coordinating multiple curators and multiple curation topics; and 2) an integrated curation page for various dataset information and automatic keyword highlighting helps curators improve their efficiency. By improving the productivity of manual curation, GEOMetaCuration is expected to benefit the biocuration community and to contribute to computational generation of biological insights using large-scale biological data.

Materials and methods

System implementation

GEOMetaCuration is designed to facilitate the manual curation process of GEO metadata, including the following steps: 1) downloading from GEO the metadata of an initial list of datasets for a specific curation topic; 2) assigning the metadata of each dataset to multiple curators; 3) (curators) curating the metadata by examining relevant materials including dataset descriptions, sample descriptions, abstracts of corresponding papers, and sometimes full papers; and 4) collecting intermediate results from curators and finalizing them (Figure 1). Initially, the GEO dataset metadata for a curation topic are downloaded by E-Utilities (9) and are imported to a database in GEOMetaCuration. On the GEOMetaCuration administration webpage, the metadata are assigned to curators. To verify curation consistency, the metadata of each dataset are curated by multiple curators. Curators curate their assigned datasets independently using the web interface, and the results are stored in the database. GEOMetaCuration then automatically collects intermediate curation results for all datasets. Results are cross-checked for integration, and only consistent results are recorded. Inconsistency in the remaining intermediate results is resolved by the administrator.

GEOMetaCuration is implemented in Django (https://www.djangoproject.com), a high-level Python web framework. The MySQL database is used for data storage. Figure 2 shows the database schema to support multiple curation topics at the same time. To facilitate curating each dataset, a novel design feature combines three types of information on an integrated curation panel for each dataset: the dataset description from GEO, the abstract of the related paper, and the full text. To speed up the recognition of useful information, predefined keywords are highlighted on the panel. GEOMetaCuration is portable and can be customized by other researchers for other curation tasks, such as literature curation. To make it freely accessible, GEOMetaCuration is published under the GNU General Public License V3. Its source code can be freely downloaded at https://bitbucket.com/yubiolab/GEOMetaCuration.

Results

Features and interface of GEOMetaCuration

To facilitate manual curation of GEO metadata, GEOMetaCuration has a feature to automatically load and cache related webpages, including GEO dataset description webpages, PubMed abstract webpages, and PubMed Central (PMC) full text webpages. Figure 3 shows the preloaded webpages for the GEO dataset GSE10192. The default page of a dataset is the GEO description page, which includes general information such as the experimental design, number of samples, and measurement platform. For example, Figure 3a shows the GEO dataset description page for GSE10192, a mouse microarray dataset with the treatment of rosiglitazone. Figure 3b shows the preloaded PubMed abstract webpage for the corresponding

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024

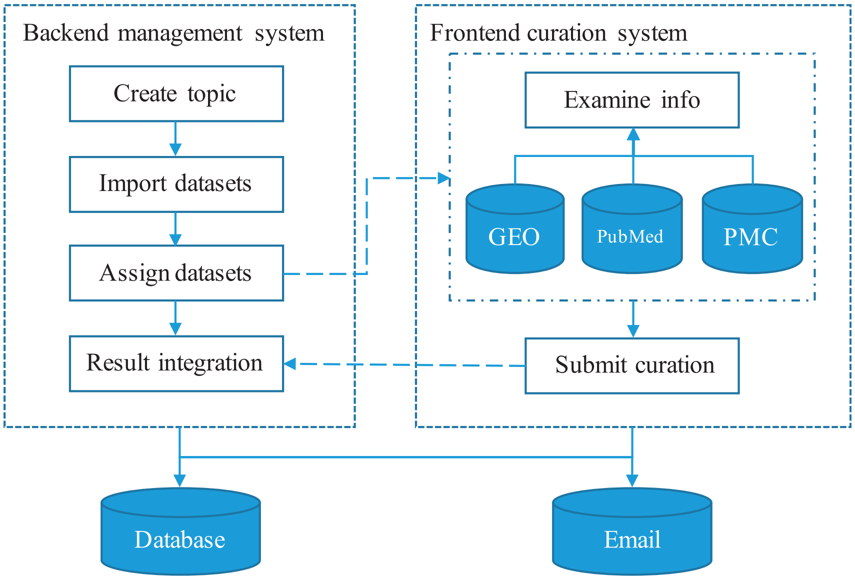


Figure 1. Workflow of GEOMetaCuration. 1) The administrator creates a topic and imports the metadata of a list of datasets to GEOMetaCuration. 2) These datasets are assigned to multiple curators to ensure curation accuracy. 3) The system sends the curators notification emails about the assignments. 4) The curators curate the assigned datasets independently. 5) The curated results are recorded on GEOMetaCuration. 6) The administrator collects the curation results from GEOMetaCuration.

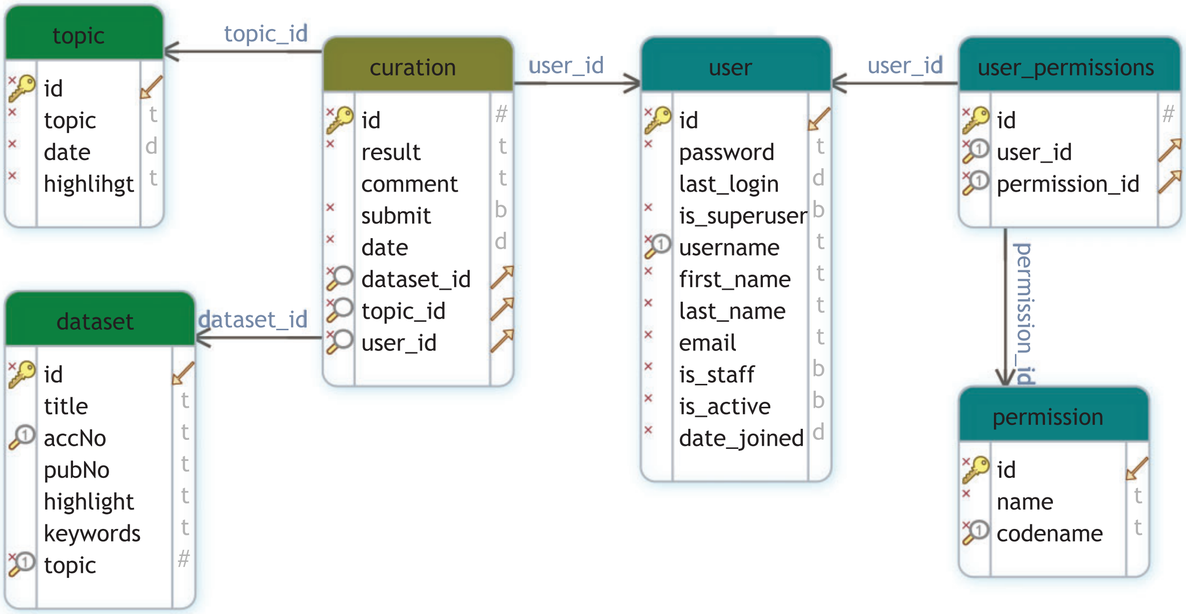
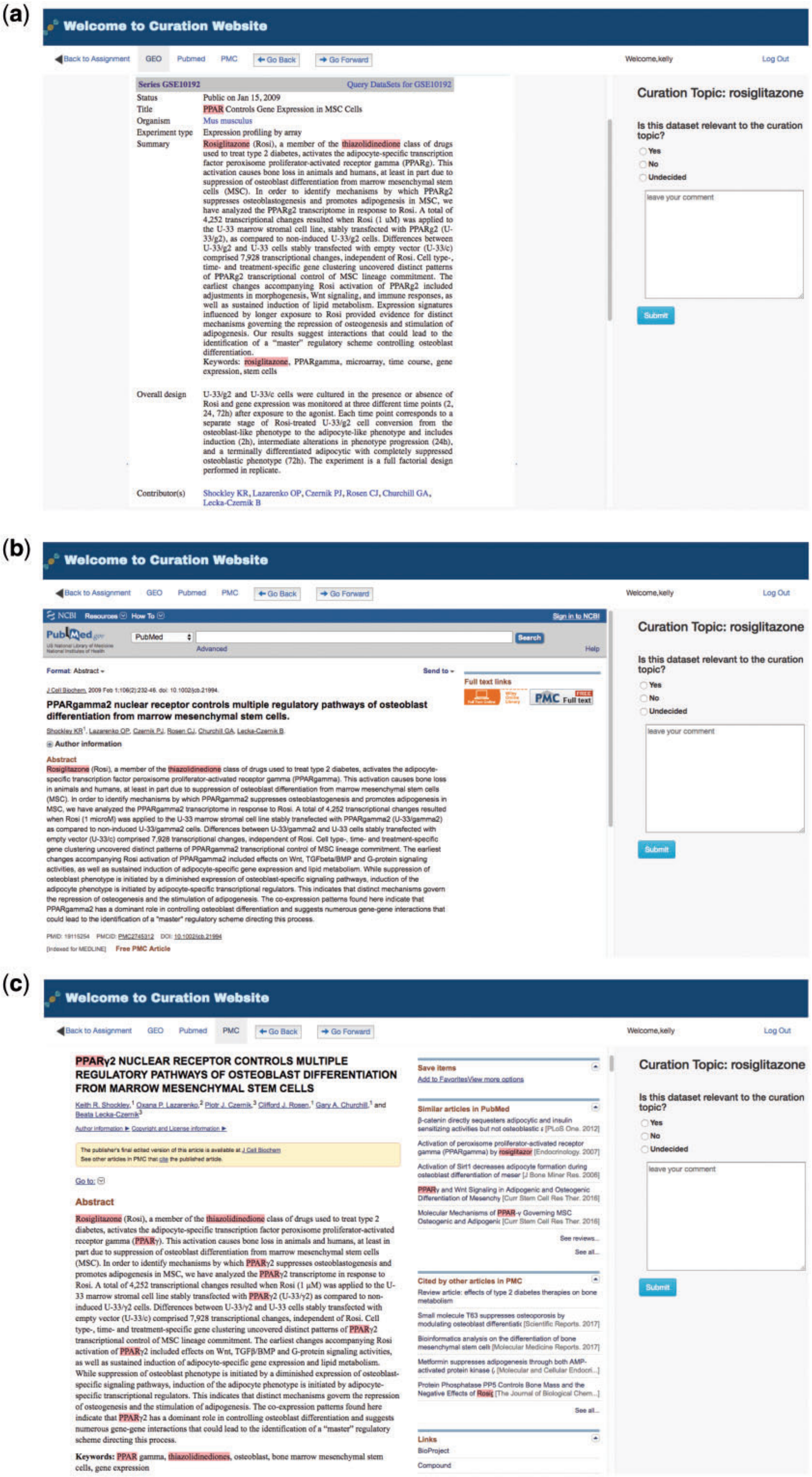


Figure 2. Database schema for GEOMetaCuration. This schema diagram contains six tables that can be classified into three groups: *user management* (user, permission, user_permission), *dataset* (dataset and topic), and *curation*, which contains the curation result of each dataset by each curator.

reference of this dataset, which can allow curators to inspect more-detailed information. Figure 3c shows the corresponding full text webpage from PMC, delivering the complete experimental information for ultimate curation decisions. GEOMetaCuration also automatically caches the webpages to save time spent repeatedly requesting the same GEO, PubMed, and PMC webpages. The preloading and caching features can expedite the curation process.

To speed up the recognition of key information needed for curation, GEOMetaCuration includes a feature to highlight predefined keywords automatically. For example, Figure 3 shows highlighted keywords 'rosiglitazone,' 'PPAR,' and 'thiazolidinedione' for dataset GSE10192. Using these visual cues, curators can look through webpages quickly in order to make curation decisions.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024



(a)

Series GSE10192	Query DataSets for GSE10192
Status	Public on Jan 15, 2009
Title	PPAR Controls Gene Expression in MSC Cells
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	Rosiglitazone (Rosi), a member of the thiazolidinedione class of drugs used to treat type 2 diabetes, activates the adipocyte-specific transcription factor peroxisome proliferator-activated receptor gamma (PPARg). This activation causes bone loss in animals and humans, at least in part due to suppression of osteoblast differentiation from marrow mesenchymal stem cells (MSC). In order to identify mechanisms by which PPARg2 suppresses osteoblastogenesis and promotes adipogenesis in MSC, we have analyzed the PPARg2 transcriptome in response to Rosi. A total of 4,252 transcriptional changes resulted when Rosi (1 uM) was applied to the U-33 marrow stromal cell line, stably transfected with PPARg2 (U-33/g2), as compared to non-induced U-33/g2 cells. Differences between U-33/g2 and U-33 cells stably transfected with empty vector (U-33/c) comprised 7,928 transcriptional changes, independent of Rosi. Cell type-, time- and treatment-specific gene clustering uncovered distinct patterns of PPARg2 transcriptional control of MSC lineage commitment. The earliest changes accompanying Rosi activation of PPARg2 included adjustments in morphogenesis, Wnt signaling, and immune responses, as well as sustained induction of lipid metabolism. Expression signatures influenced by longer exposure to Rosi provided evidence for distinct mechanisms governing the repression of osteogenesis and stimulation of adipogenesis. Our results suggest interactions that could lead to the identification of a "master" regulatory scheme controlling osteoblast differentiation. Keywords: rosiglitazone, PPARgamma, microarray, time course, gene expression, stem cells
Overall design	U-33/g2 and U-33/c cells were cultured in the presence or absence of Rosi and gene expression was monitored at three different time points (2, 24, 72h) after exposure to the agonist. Each time point corresponds to a separate stage of Rosi-treated U-33/g2 cell conversion from the osteoblast-like phenotype to the adipocyte-like phenotype and includes induction (2h), intermediate alterations in phenotype progression (24h), and a terminally differentiated adipocytic with completely suppressed osteoblastic phenotype (72h). The experiment is a full factorial design performed in replicate.
Contributor(s)	Shockley KR, Lazarenko OP, Czernik PJ, Rosen CJ, Churchill GA, Lecka-Czernik B

Curation Topic: rosiglitazone

Is this dataset relevant to the curation topic?

○ Yes
○ No
○ Undecided

leave your comment

(b)

J Cell Biochem. 2009 Feb 1;106(2):232-46. doi: 10.1002/jcb.21994.

PPARgamma2 nuclear receptor controls multiple regulatory pathways of osteoblast differentiation from marrow mesenchymal stem cells.

Shockley KR[1], Lazarenko OP, Czernik PJ, Rosen CJ, Churchill GA, Lecka-Czernik B.

⊕ Author information

Abstract

Rosiglitazone (Rosi), a member of the thiazolidinedione class of drugs used to treat type 2 diabetes, activates the adipocyte-specific transcription factor peroxisome proliferator-activated receptor gamma (PPARgamma). This activation causes bone loss in animals and humans, at least in part due to suppression of osteoblast differentiation from marrow mesenchymal stem cells (MSC). In order to identify mechanisms by which PPARgamma2 suppresses osteoblastogenesis and promotes adipogenesis in MSC, we have analyzed the PPARgamma2 transcriptome in response to Rosi. A total of 4,252 transcriptional changes resulted when Rosi (1 microM) was applied to the U-33 marrow stromal cell line stably transfected with PPARgamma2 (U-33/gamma2) as compared to non-induced U-33/gamma2 cells. Differences between U-33/gamma2 and U-33 cells stably transfected with empty vector (U-33/c) comprised 7,928 transcriptional changes, independent of Rosi. Cell type-, time- and treatment-specific gene clustering uncovered distinct patterns of PPARgamma2 transcriptional control of MSC lineage commitment. The earliest changes accompanying Rosi activation of PPARgamma2 included effects on Wnt, TGFbeta/BMP and G-protein signaling activities, as well as sustained induction of adipocyte-specific gene expression and lipid metabolism. While suppression of osteoblast phenotype is initiated by a diminished expression of osteoblast-specific signaling pathways, induction of the adipocyte phenotype is initiated by adipocyte-specific transcriptional regulators. This indicates that distinct mechanisms govern the repression of osteogenesis and the stimulation of adipogenesis. The co-expression patterns found here indicate that PPARgamma2 has a dominant role in controlling osteoblast differentiation and suggests numerous gene-gene interactions that could lead to the identification of a "master" regulatory scheme directing this process.

PMID: 19115254 PMCID: PMC2745312 DOI: 10.1002/jcb.21994

[Indexed for MEDLINE] Free PMC Article

Curation Topic: rosiglitazone

Is this dataset relevant to the curation topic?

○ Yes
○ No
○ Undecided

leave your comment

(c)

PPARγ2 NUCLEAR RECEPTOR CONTROLS MULTIPLE REGULATORY PATHWAYS OF OSTEOBLAST DIFFERENTIATION FROM MARROW MESENCHYMAL STEM CELLS

Keith R. Shockley,[1] Oxana P. Lazarenko,[2] Piotr J. Czernik,[3] Clifford J. Rosen,[1] Gary A. Churchill,[1] and Beata Lecka-Czernik[3]

Author information ► Copyright and License information ►

The publisher's final edited version of this article is available at J Cell Biochem
See other articles in PMC that cite the published article.

Go to: ⊡

Abstract

Rosiglitazone (Rosi), a member of the thiazolidinedione class of drugs used to treat type 2 diabetes, activates the adipocyte-specific transcription factor peroxisome proliferator-activated receptor gamma (PPARγ). This activation causes bone loss in animals and humans, at least in part due to suppression of osteoblast differentiation from marrow mesenchymal stem cells (MSC). In order to identify mechanisms by which PPARγ2 suppresses osteoblastogenesis and promotes adipogenesis in MSC, we have analyzed the PPARγ2 transcriptome in response to Rosi. A total of 4,252 transcriptional changes resulted when Rosi (1 μM) was applied to the U-33 marrow stromal cell line stably transfected with PPARγ2 (U-33/γ2) as compared to non-induced U-33/γ2 cells. Differences between U-33/γ2 and U-33 cells stably transfected with empty vector (U-33/c) comprised 7,928 transcriptional changes, independent of Rosi. Cell type-, time- and treatment-specific gene clustering uncovered distinct patterns of PPARγ2 transcriptional control of MSC lineage commitment. The earliest changes accompanying Rosi activation of PPARγ2 included effects on Wnt, TGFβ/BMP and G-protein signaling activities, as well as sustained induction of adipocyte-specific gene expression and lipid metabolism. While suppression of osteoblast phenotype is initiated by a diminished expression of osteoblast-specific signaling pathways, induction of the adipocyte phenotype is initiated by adipocyte-specific transcriptional regulators. This indicates that distinct mechanisms govern the repression of osteogenesis and the stimulation of adipogenesis. The co-expression patterns found here indicate that PPARγ2 has a dominant role in controlling osteoblast differentiation and suggests numerous gene-gene interactions that could lead to the identification of a "master" regulatory scheme directing this process.

Keywords: PPAR gamma, thiazolidinediones, osteoblast, bone marrow mesenchymal stem cells, gene expression

Save items
Add to FavoritesView more options

Similar articles in PubMed
- β-catenin directly sequesters adipocytic and insulin sensitizing activities but not osteoblastic ε [PLoS One. 2012]
- Activation of peroxisome proliferator-activated receptor gamma (PPARgamma) by rosiglitazor [Endocrinology. 2007]
- Activation of Sirt1 decreases adipocyte formation during osteoblast differentiation of meser [J Bone Miner Res. 2006]
- PPARγ and Wnt Signaling in Adipogenic and Osteogenic Differentiation of Mesenchy [Curr Stem Cell Res Ther. 2016]
- Molecular Mechanisms of PPAR-γ Governing MSC Osteogenic and Adipogenic [Curr Stem Cell Res Ther. 2016]

See reviews...
See all...

Cited by other articles in PMC
- Review article: effects of type 2 diabetes therapies on bone metabolism
- Small molecule T63 suppresses osteoporosis by modulating osteoblast differentiati [Scientific Reports. 2017]
- Bioinformatics analysis on the differentiation of bone mesenchymal stem cell [Molecular Medicine Reports. 2017]
- Metformin suppresses adipogenesis through both AMP-activated protein kinase (, [Molecular and Cellular Endocri...]
- Protein Phosphatase PP5 Controls Bone Mass and the Negative Effects of Rosig [The Journal of Biological Chem...]

See all...

Links
- BioProject
- Compound

Curation Topic: rosiglitazone

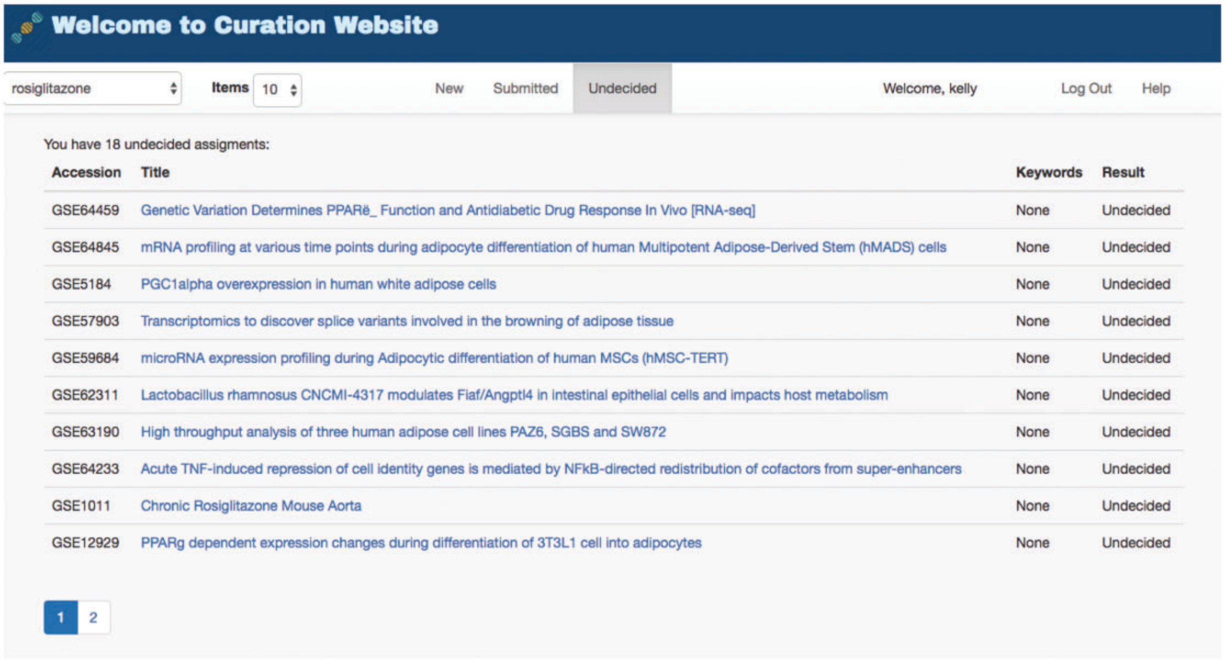
Is this dataset relevant to the curation topic?

○ Yes
○ No
○ Undecided

leave your comment

Figure 3. Preloaded GEOMetaCuration webpages for the metadata curation of GSE10192. To expedite metadata curation, GEOMetaCuration is designed to preload webpages for curators to avoid waiting for these webpages to open. To assist in quickly identifying relevant information, predefined keywords are highlighted on the webpages. For the example of GSE10192, keywords are highlighted on three preloaded webpages: (**a**) the GEO dataset description webpage, (**b**) the PubMed abstract webpage of the corresponding publication, and (**c**) the PMC webpage for the full paper.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024



You have 18 undecided assigments:

Accession	Title	Keywords	Result
GSE64459	Genetic Variation Determines PPARë_ Function and Antidiabetic Drug Response In Vivo [RNA-seq]	None	Undecided
GSE64845	mRNA profiling at various time points during adipocyte differentiation of human Multipotent Adipose-Derived Stem (hMADS) cells	None	Undecided
GSE5184	PGC1alpha overexpression in human white adipose cells	None	Undecided
GSE57903	Transcriptomics to discover splice variants involved in the browning of adipose tissue	None	Undecided
GSE59684	microRNA expression profiling during Adipocytic differentiation of human MSCs (hMSC-TERT)	None	Undecided
GSE62311	Lactobacillus rhamnosus CNCMI-4317 modulates Fiaf/Angptl4 in intestinal epithelial cells and impacts host metabolism	None	Undecided
GSE63190	High throughput analysis of three human adipose cell lines PAZ6, SGBS and SW872	None	Undecided
GSE64233	Acute TNF-induced repression of cell identity genes is mediated by NFkB-directed redistribution of cofactors from super-enhancers	None	Undecided
GSE1011	Chronic Rosiglitazone Mouse Aorta	None	Undecided
GSE12929	PPARg dependent expression changes during differentiation of 3T3L1 cell into adipocytes	None	Undecided

Figure 4. Curation task management webpage for a curator. To facilitate the management of datasets for curation, GEOMetaCuration has three tabs for assigned datasets. The 'New' tab shows datasets to be curated. The 'Submitted' tab shows datasets that have been curated. The 'Undecided' tab shows datasets whose curation results have not been determined yet. Curation topics can be selected on the drop-down menu at the top-left corner ('rosiglitazone' is currently selected in this example). The hyperlink of each dataset links to the corresponding detailed curation webpage, such as the one shown in Figure 3.

To help manage the complete lifecycle of curation, GEOMetaCuration can group curation tasks into three stages: 'New,' 'Submitted' and 'Undecided,' shown as three tabs in Figure 4. The 'New' tab lists the datasets to be curated. The 'Submitted' tab lists the datasets that have been curated with a confident decision from the curator. The 'Undecided' tab lists the datasets on which the curator has started but has not reached a final decision (usually because the corresponding metadata are complicated and need more detailed examination). These tabs facilitate efficient organization of datasets assigned to curators.

To manage the curation task efficiently, GEOMetaCuration fulfills an administration service for different curation tasks. For example, Figure 5 shows assignment of the curation task for rosiglitazone. The curators and their assigned datasets can be easily selected on the administration webpage. Datasets can be easily assigned to multiple curators at once. This administration service enables powerful management of many curation tasks.

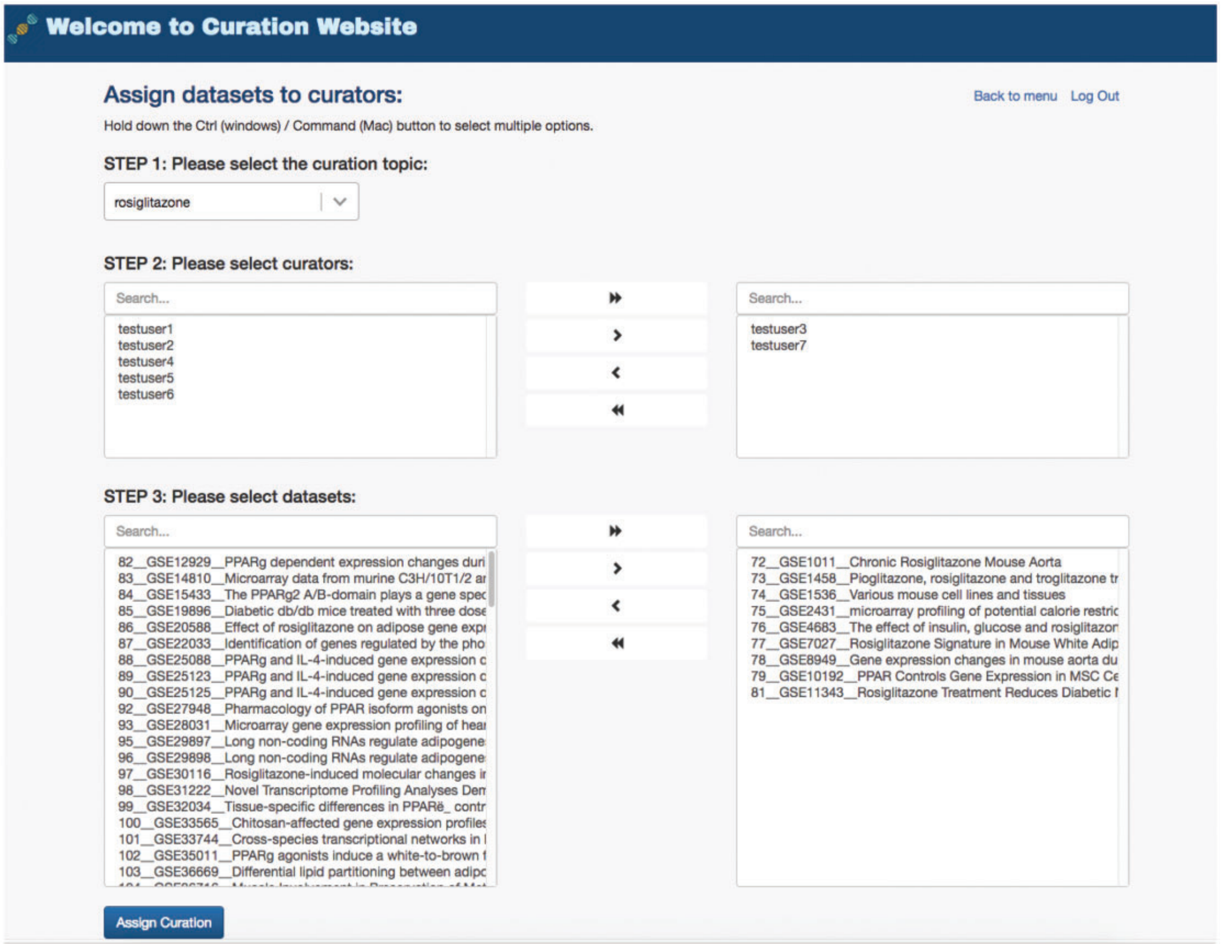
Improved curation time with GEOMetaCuration

We designed the following experiment to confirm improvement to the curation process using GEOMetaCuration. A typical curation process includes two components: 1) performing mechanical operations such as opening the webpages related to each dataset; and 2) looking through the webpages and make a curation decision. Because our tool is designed to only improve the time of the first component and because the time of the second component varies for different curation tasks and for different curators, only the time of the first component was measured and compared.

Google Drive was used in the baseline curation process. Time was recorded for the following activities: 1) copying GEO dataset accession numbers from Google Drive (which contains the GEO gene expression datasets about 'rosiglitazone'); 2) Googling the dataset accession numbers and clicking the returned hyperlinks to open the GEO dataset description webpages; 3) clicking the PubMed URLs on these description webpages; 4) clicking the PMC manuscript URLs on the PubMed webpages; 5) typing Ctrl-F and searching for keywords (rosiglitazone, PPAR, and thiazolidinedione) in each webpage opened in steps 2 to 4; and 6) switching back to Google Drive and putting mock results into the rows corresponding to the curated datasets.

Using our curation website, time was recorded for clicking dataset webpages and typing mock results. The time of both with and without the caching feature was recorded to confirm the improvement experience using this feature. Figure 6 shows the time for each of the first 20 gene expression datasets in Supplementary Table S2. It demonstrates that GEOMetaCuration reduces the time for curation by an average of 60 seconds each time when the information of a dataset is viewed, and the caching

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024



Welcome to Curation Website

Assign datasets to curators:

Hold down the Ctrl (windows) / Command (Mac) button to select multiple options.

STEP 1: Please select the curation topic:

rosiglitazone

STEP 2: Please select curators:

testuser1
testuser2
testuser4
testuser5
testuser6

testuser3
testuser7

STEP 3: Please select datasets:

82__GSE12929__PPARg dependent expression changes duri
83__GSE14810__Microarray data from murine C3H/10T1/2 ar
84__GSE15433__The PPARg2 A/B-domain plays a gene spec
85__GSE19896__Diabetic db/db mice treated with three dose
86__GSE20588__Effect of rosiglitazone on adipose gene expr
87__GSE22033__Identification of genes regulated by the pho
88__GSE25088__PPARg and IL-4-induced gene expression c
89__GSE25123__PPARg and IL-4-induced gene expression c
90__GSE25125__PPARg and IL-4-induced gene expression c
92__GSE27948__Pharmacology of PPAR isoform agonists on
93__GSE28031__Microarray gene expression profiling of hear
95__GSE29897__Long non-coding RNAs regulate adipogene
96__GSE29898__Long non-coding RNAs regulate adipogene
97__GSE30116__Rosiglitazone-induced molecular changes ir
98__GSE31222__Novel Transcriptome Profiling Analyses Den
99__GSE32034__Tissue-specific differences in PPARë_ contr
100__GSE33565__Chitosan-affected gene expression profiles
101__GSE33744__Cross-species transcriptional networks in l
102__GSE35011__PPARg agonists induce a white-to-brown f
103__GSE36669__Differential lipid partitioning between adipc

72__GSE1011__Chronic Rosiglitazone Mouse Aorta
73__GSE1458__Pioglitazone, rosiglitazone and troglitazone tr
74__GSE1536__Various mouse cell lines and tissues
75__GSE2431__microarray profiling of potential calorie restric
76__GSE4683__The effect of insulin, glucose and rosiglitazon
77__GSE7027__Rosiglitazone Signature in Mouse White Adip
78__GSE8949__Gene expression changes in mouse aorta du
79__GSE10192__PPAR Controls Gene Expression in MSC Ce
81__GSE11343__Rosiglitazone Treatment Reduces Diabetic I

Figure 5. Administration webpage for assigning datasets to multiple curators. The administration webpage allows efficient assignment of datasets to multiple curators for any given specific curation task. Once s/he selects a curation topic, the administrator can then concurrently select the curators and the datasets for an assignment.

feature further reduces it. By reducing this time, the chance for curator distraction is decreased, which can also improve curation efficiency.

Use case—the curation of human gene expression datasets with rosiglitazone as a treatment

To show the features of GEOMetaCuration, we created a demo website (http://geometacuration.yubiolab.org) using the curation of datasets related to rosiglitazone. Rosiglitazone enhances thermogenesis in brown adipose tissues, which can be helpful to fight obesity and reduce the chances of developing type 2 diabetes (10, 11). To study its function in detail, it is critical to collect the complete gene expression data with rosiglitazone as a treatment. The metadata of datasets mentioning 'rosiglitazone' were retrieved by querying the database GEO DataSets (https://www.ncbi.nlm.nih.gov/gds) via E-utilities (9) using the following term:

> (rosiglitazone[Description] OR rosiglitazone[Title]) AND ("*Homo sapiens*"[porgn:__txid9606]) AND gse[Entry Type] AND (("expression profiling by array"[DataSet Type] OR "non-coding rna profiling by array"[DataSet Type]) OR ("expression profiling by high throughput sequencing"[DataSet Type] OR "non-coding rna profiling by high throughput sequencing"[DataSet Type])) NOT SuperSeries[All Fields].

The retrieved metadata were first loaded into the MySQL database on the demo website. Then, these datasets were assigned to each curator by the website administrator (Figure 4). A notification email was then sent to each curator about the assignment. Each dataset was assigned to multiple curators, and the administrator compared inconsistent answers. To help the curator rapidly locate important information about rosiglitazone, the keywords 'rosiglitazone,' 'PPAR,' and 'thiazolidinedione' were loaded into the back-end management system and were highlighted with different colors on the front end. By examining the metadata of a large number of datasets (Supplementary Table S2), the final curation results were collected (Supplementary Table S1). This use case confirms the utility of GEOMetaCuration.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024

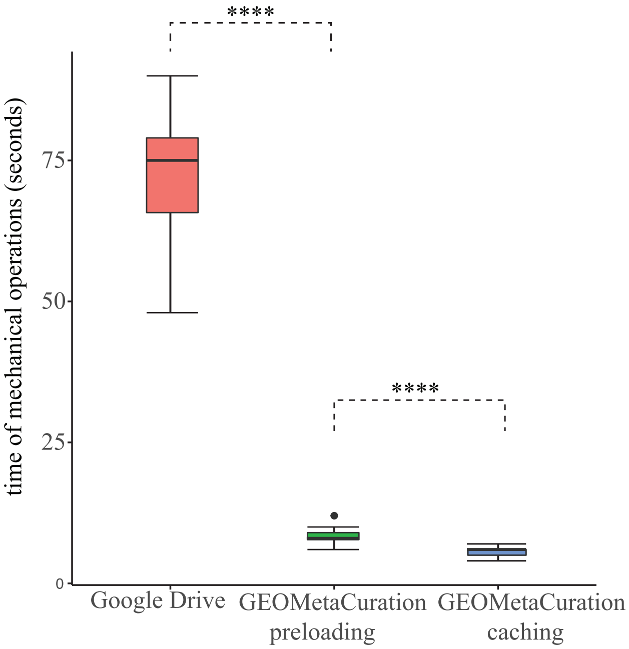


Figure 6. Time spent on mechanical operations required for curation of each of 20 datasets. GEOMetaCuration significantly reduces curation time by eliminating mechanical operations like opening webpages, searching important information, and typing in results. The webpage caching feature further reduces curation time by avoiding repetitive requests to GEO, PubMed, and PMC. **** indicates $P < 0.0001$ (one-sided *t*-test).

Discussion

Accurate metadata curation is imperative for integrating large-scale biological data to generate biological insights in biomedical research. Since the data for a specific biomedical research problem are usually limited, missing known datasets may result in unnecessary duplication of experiments, leading to increased time and cost and the reduction of research productivity. Meanwhile, although more than 100 000 datasets and more than 2 million samples are hosted on GEO, the number of datasets for a certain biological problem is usually still limited. For example, at the time of writing, a query of general keywords such as 'diabetes,' 'heart disease,' and 'Alzheimer's disease' in GEO retrieved ∼400, ∼400 and ∼100 human gene expression datasets, respectively. For this limited number of datasets, it would be feasible to manually curate the metadata in just a couple of weeks by well-trained curators. For this reason, we implemented GEOMetaCuration to facilitate the curation process.

Conventional metadata curation is prolonged and inefficient. For example, to construct a metadata database such as RNASeqMetaDB and SFMetaDB, the metadata must be split into multiple files and shared with different curators. Curators then need to manually check the GEO webpages and associated publications. Upon finishing the curation tasks, they submit their intermediate results, which are collected and integrated to construct the final results. It is burdensome to make their notations uniform since different curators have personal habits for marking answers and taking notes. To address these problems, we designed various critical features for metadata curation in GEOMetaCuration. For example, the webpage preloading feature saves curation time by avoiding manually opening webpages related to curation. The keyword highlight feature expedites the identification of relevant information in metadata. Furthermore, the administration service streamlines the management of curators and their tasks.

GEOMetaCuration can enable productive metadata curation of GEO datasets. As its name suggests, GEOMetaCuration currently only supports GEO metadata. But because it is based on the industry-tested web development framework of Django/Python, GEOMetaCuration can be customized for other curation applications, such as literature curation. Its code is open-sourced and can be modified by other researchers for their curation tasks. Therefore, GEOMetaCuration is a useful biocuration application that will eventually contribute to the generation of biological insights using large-scale biological datasets.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgement

The authors thank Bing Jiang for her contribution to GEOMetaCuration.

Funding

This work was supported by startup funding to P.Y. from the ECE department and Texas A&M Engineering Experiment Station/ Dwight Look College of Engineering at Texas A&M University, by funding from TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE) at Texas A&M University, by TEES seed grant, and by Texas A&M University-CAPES Research Grant Program. The open access publishing fees for this article have been covered by the Texas A&M University Open Access to Knowledge Fund (OAKFund), supported by the University Libraries and the Office of the Vice President for Research.

Conflict of interest. None declared.

References

1. Dauga,D. (2015) Biocuration: a new challenge for the tunicate community. *Genesis*, **53**, 132–142.
2. Cahan,P., Li,H., Morris,S.A. *et al.* (2014) CellNet: network biology applied to stem cell engineering. *Cell*, **158**, 903–915.
3. Howe,D., Costanzo,M., Fey,P. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024

4. Barrett,T., Wilhite,S.E., Ledoux,P. *et al.* (2013) NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.*, **41**, D991–D995.
5. Bernstein,M.N., Doan,A. and Dewey,C.N. (2017) MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, **33**, 2914–2923.
6. Kilicoglu,H. (2017) Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Brief Bioinform.*, doi: 10.1093/bib/bbx057.
7. Guo,Z.Y., Tzvetkova,B., Bassik,J.M. *et al.* (2015) RNASeqMetaDB: a database and web server for navigating metadata of publicly available mouse RNA-Seq datasets. *Bioinformatics*, **31**, 4038–4040.
8. Li,J., Tseng,C.S., Federico,A. *et al.* (2017) SFMetaDB: a comprehensive annotation of mouse RNA splicing factor RNA-Seq datasets. *Database*, **2017**, 1–7.
9. Agarwala,R., Barrett,T., Beck,J. *et al.* (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
10. Li,Y.L., Li,X., Jiang,T.T. *et al.* (2017) An additive effect of promoting thermogenic gene expression in mice adipose-derived stromal vascular cells by combination of rosiglitazone and CL316, 243. *Int. J. Mol. Sci.*, **18**, 1002.
11. Ohno,H., Shinoda,K., Spiegelman,B.M. *et al.* (2012) PPAR gamma agonists induce a white-to-brown fat conversion through stabilization of PRDM16 Protein. *Cell Metab.*, **15**, 395–404.

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bay019/4953404 by guest on 21 May 2024