Biophysical Reviews (2019) 11:103–110
https://doi.org/10.1007/s12551-018-0490-8

REVIEW

Mining data and metadata from the gene expression omnibus

Zichen Wang[1] · Alexander Lachmann[1] · Avi Ma'ayan[1]

Received: 24 October 2018 / Accepted: 4 December 2018 / Published online: 29 December 2018

Abstract

Publicly available gene expression datasets deposited in the Gene Expression Omnibus (GEO) are growing at an accelerating rate. Such datasets hold great value for knowledge discovery, particularly when integrated. Although numerous software platforms and tools have been developed to enable reanalysis and integration of individual, or groups, of GEO datasets, large-scale reuse of those datasets is impeded by minimal requirements for standardized metadata both at the study and sample levels as well as uniform processing of the data across studies. Here, we review methodologies developed to facilitate the systematic curation and processing of publicly available gene expression datasets from GEO. We identify trends for advanced metadata curation and summarize approaches for reprocessing the data within the entire GEO repository.

Keywords GEO · Gene Expression Omnibus · Computational data curation · Natural language processing · FAIR principles

Abbreviations

NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
SRA	Sequence Read Archive
DE	Differential expression
DEG	Differentially expressed genes
NCBO	National Center for Biomedical Ontology
MOOC	Massive open online course
ML	Machine learning
NLP	Natural language processing
NER	Named-entity recognition
CEDAR	Center for Expanded Data Annotation and Retrieval
LSTM	Long short-term memory
CNN	Convolutional neural network
CRF	Conditional random fields
AL	Active learning
PMC	PubMed Central
FAIR	Findable, accessible, interoperable and reproducible
API	Application programming interface
JSON-LD	JavaScript Object Notation for Linked Data

This article is part of a Special Issue on 'Big Data' edited by Joshua WK Ho and Eleni Giannoulatou.

✉ Zichen Wang
zichen.wang@mssm.edu

[1] BD2K-LINCS Data Coordination and Integration Center; Knowledge Management Center for the Illuminating the Druggable Genome; Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, Box 1603, One Gustave L. Levy Place, New York, NY 10029, USA

Introduction

Gene expression datasets are accumulating rapidly in public repositories such as the NCBI's Gene Expression Omnibus (GEO) (Barrett et al. 2013) and the Sequence Read Archive (SRA) (Kodama et al. 2012) as well as ArrayExpress (Rustici et al. 2013). That is partly driven by the emergence of new and improved transcriptomic profiling technologies such as RNA sequencing (RNA-seq) (Fig. 1). In addition, most journals now mandate the deposition of transcriptomics data as a requirement for publication, with the goal of enabling reproducibility and data reuse. Reanalysis and integration of themed collections of gene expression datasets can produce new insights into the underlying biological mechanisms under investigation. For instance, meta-analysis of multiple datasets for a disease can help in discovering the most consistently differentially expressed genes (DEGs) and the pathways that these genes belong. In addition, consistent DEGs can become biomarkers and drug targets. Similarly, curated collections of gene expression signatures can serve as a Connectivity Map reference database for matching user-submitted signatures of DEGs with annotated and curated signatures (Lamb et al. 2006; Subramanian et al. 2017). Similarly, curated signatures can be converted to gene set libraries for gene

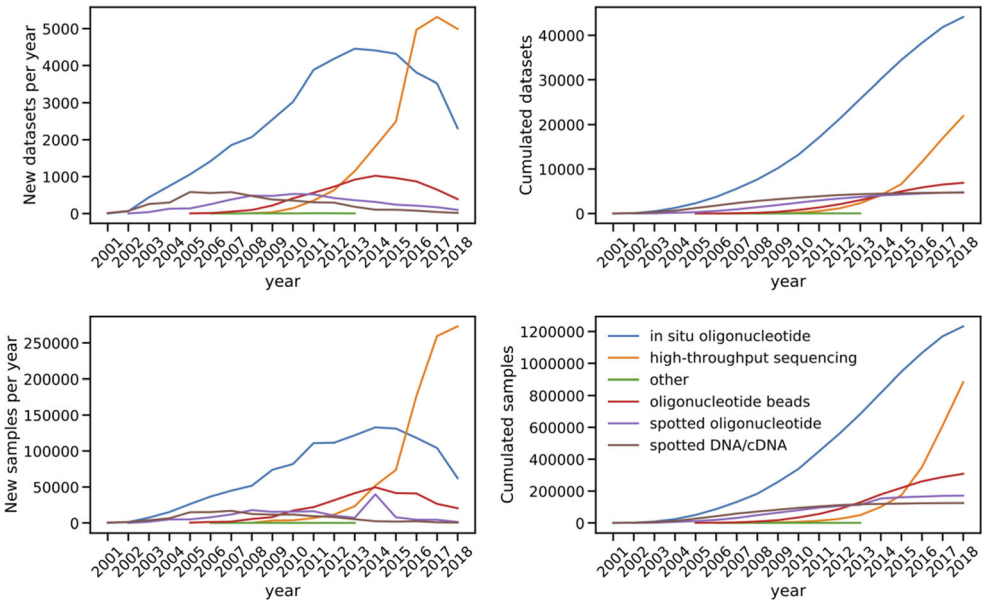

Fig. 1 The growth of publicly available gene expression datasets and samples from GEO over time. Plots on the top panel show the growth of gene expression datasets from different transcriptomic profiling technologies over time, whereas plots on the bottom panel show the growth of individual samples from those datasets. The plots were made on September 2018. Hence, the total for 2018 cover only part of the year

set enrichment analyses (Chen et al. 2013; Kuleshov et al. 2016; Subramanian et al. 2005). In addition, curated signatures can be compared for reproducibility across multiple independent studies (Gundersen et al. 2016), or for finding unexpected relationships between drugs, genes, and diseases (Wang et al. 2016; Chen and Butte 2016; Cheng et al. 2014).

Several software tools have been developed for reanalyzing individual or collections of datasets from GEO (Table 1). Those tools enable users to search GEO for relevant studies and then retrieve specific datasets for further analysis. In addition to those tools, approaches have been developed to uniformly reprocess all the microarray or RNA-seq datasets in GEO. The uniformly reprocessed gene expression datasets can be organized into databases that serve as search engines that enable knowledge discovery at the data level. Prominent examples include ExpressionBlast (Zinman et al. 2013), Recount2 (Collado-Torres et al. 2017), ARCHS4 (Lachmann et al. 2018), and SEEK (Zhu et al. 2015). These resources processed a large number of microarray and RNA-seq samples to build search engines for gene expression profiles and co-expression modules. Recent advances in cloud computing infrastructure, efficient cloud-enabled aligners such as Rail-RNA (Nellore et al. 2017), and alignment-free RNA-seq quantification methods such as Kallisto (Bray et al. 2016) enable the large-scale uniform reprocessing of RNA-seq datasets from GEO. Such efforts include Recount2 (Collado-Torres et al. 2017) and ARCHS4

Table 1 Software tools developed for reanalyzing and further annotating GEO datasets

Tool	Citation	Individual/ multiple	Type	Note	Limitations
GEO2R	(Barrett et al. 2013)	Individual	Web	Implements GUI that generate graphs and R script	Limited graphical visualizations; only implements DE analysis; limited to microarray data
shinyGEO	(Dumas et al. 2016)	Individual	Web	R Shiny extension of GEO2R with improved graphics	DE analysis only available for individual genes; limited to microarray data
GEOquery	(Davis and Meltzer 2007)	Individual	R package	Bridge between GEO and BioConductor to enable analyses of GEO datasets in various BioConductor packages	Requires users to be proficient in R and Bioconductor packages; limited to microarray data
GEO2Enrichr	(Gundersen et al. 2015)	Individual	Brower extension	Identifies DEGs and pipe to enrichment analysis tool	Limited to microarray data; limited analysis components
BioJupies	(Torre et al. 2018)	Individual	Web	Generates interactive Jupyter notebooks from RNA-seq datasets	Limited to RNA-seq data. Only allows 2 group comparison
ScanGEO	(Koeppen et al. 2017)	Multiple	Web	Identifies DEGs across multiple GEO studies matching user-specified criteria	Limited to curated GEO datasets (GDS); only supports DE analysis
ImaGEO	(Toro-Domínguez et al. 2018)	Multiple	Web	Performs nine types of meta-analysis across multiple GEO studies	Limited to microarray datasets
GEOracle	(Djordjevic et al. 2017)	Multiple	Web	Uses text mining of the GEO metadata to automatically identify perturbational GEO datasets and associated metadata	Limited to microarray datasets; only performs DE analysis

(Lachmann et al. 2018). These newer search engines provide other features besides sample search, for example, gene function prediction, average expression in tissues and cells, and systematic discovery of alternative splicing events.

However, integrating datasets across studies as well as performing meta-analyses from collections of studies is still difficult. This is mainly because of the lack of machine-readable standardized metadata at the study and sample levels. The metadata associated with gene expression studies within GEO typically do not adhere to controlled vocabularies to describe biological entities such as tissue type, cell type, cell line, gene/protein, drug/small-molecule, and disease. Instead, the authors of the datasets use semi-structured textual descriptions to annotate their study design, sample characteristics, and experimental protocols. Many GEO studies are also associated with publications indexed in PubMed, which further helps other researchers to understand the details of each study design, but does not resolve the necessity for machine-readable metadata.

Therefore, there is an urgent need for better curating and annotating publicly available gene expression datasets at scale to enable better data reuse that can facilitate new discoveries. The task of curating and annotating GEO datasets involves the identifying and mapping of biological entities such as genes/proteins, drugs/small-molecules, diseases, and cells/tissue-types at both the dataset and sample levels. Such mapping needs to be done to relevant community-accepted controlled vocabularies such as specialized ontologies available from the National Center for Biomedical Ontology (NCBO) BioPortal (Whetzel et al. 2011) and other community-accepted naming standards. Better annotation of datasets and samples will provide the basis for identifying meaningful biological contrasts among groups of samples, which can then be used for differential expression (DE) analysis. Here, we review recent advances and future perspectives in the process of curating and reprocessing publicly available gene expression datasets from GEO.

Approaches toward improving curation and annotation of GEO metadata

Multiple approaches have been developed for improving the curating of the metadata associated with publicly available studies served on the GEO repository. These methods can be broadly categorized into (1) manual curation, (2) automated natural language processing (NLP), and (3) inferring metadata directly from the gene expression profiles. In the subsequent sections, we describe recent activities within these three categories (Fig. 2).

Manual curation

Although not perfect, manual curation efforts applied to annotate GEO studies yield high-quality results. However, manual curation does not scale up to cover the tens of thousands of studies that are currently available from GEO. Since GEO, and repositories like it, are expected to drastically grow in the coming years, manual curation is in general not feasible. Crowdsourcing microtasks are projects that consist of a relatively trivial task that requires a large number of participants to complete (Good and Su 2013; Khare et al. 2015). Such an approach is one way to scale up manual metadata curation of GEO datasets. Through a massive open online course (MOOC) on Coursera, we worked together with over 70 participants from over 25 countries to identify and annotate 2460 single-gene perturbation signatures, 839 disease signatures, and 906 drug perturbation signatures from GEO (Wang et al. 2016). The collections of these signatures are served as a web portal called CRowd Extracted Expression of Differential Signatures (CREEDS). CREEDS provides the annotated signatures for query, download, and visualization. A few other similar projects were launched to curate GEO datasets using microtask crowdsourcing strategies. One such project is STARGEO, a website that facilitates the curation of GEO samples with disease phenotypes. The STARGEO project is a manual crowdsourcing curation effort that recruited graduate students to annotate samples with disease phenotypes (Hadley et al. 2017). Another similar effort called OMics Compendia Commons (OMiCC) (Shah et al. 2016) is a community-oriented framework that enables biomedical researchers to collaboratively annotate gene expression datasets and samples. OMiCC is also equipped with a web interface that lets users perform meta-analyses including differential expression analysis.

The manually curated GEO datasets facilitated the reanalysis of multiple related datasets to reveal novel biological insights. For instance, by clustering the curated signatures from genetic perturbation and diseases, we found multiple myelodysplastic syndrome (MDS) signatures from CD34+ cells that cluster with *ERBB2* overexpression signatures from MCF10A cells. Such co-clustering suggests that the upregulation of *ERBB2* and related pathways may play a role in MDS (Wang et al. 2016). Another example is the meta-analysis of inflammatory bowel disease (IBD) signatures across multiple independent studies, curated by the OMiCC platform. This analysis discovered that several peroxisome proliferator-activated receptors (PPARs) are lowly expressed in Crohn's disease (Shah et al. 2016).

While manual curation through crowdsourcing produces, in general, high-quality annotations, this approach has other drawbacks besides lack of scalability. Curators make mistakes and produce inconsistent annotations in borderline cases (Good and Su 2013; Khare et al. 2015). While this can be resolved through a double-blinded review process, having multiple curators annotate the same datasets increases the burden on the curation task many folds. For the CREEDS project, we had to spot check all entries and remove contributors that produced annotations with high error rates. Another approach

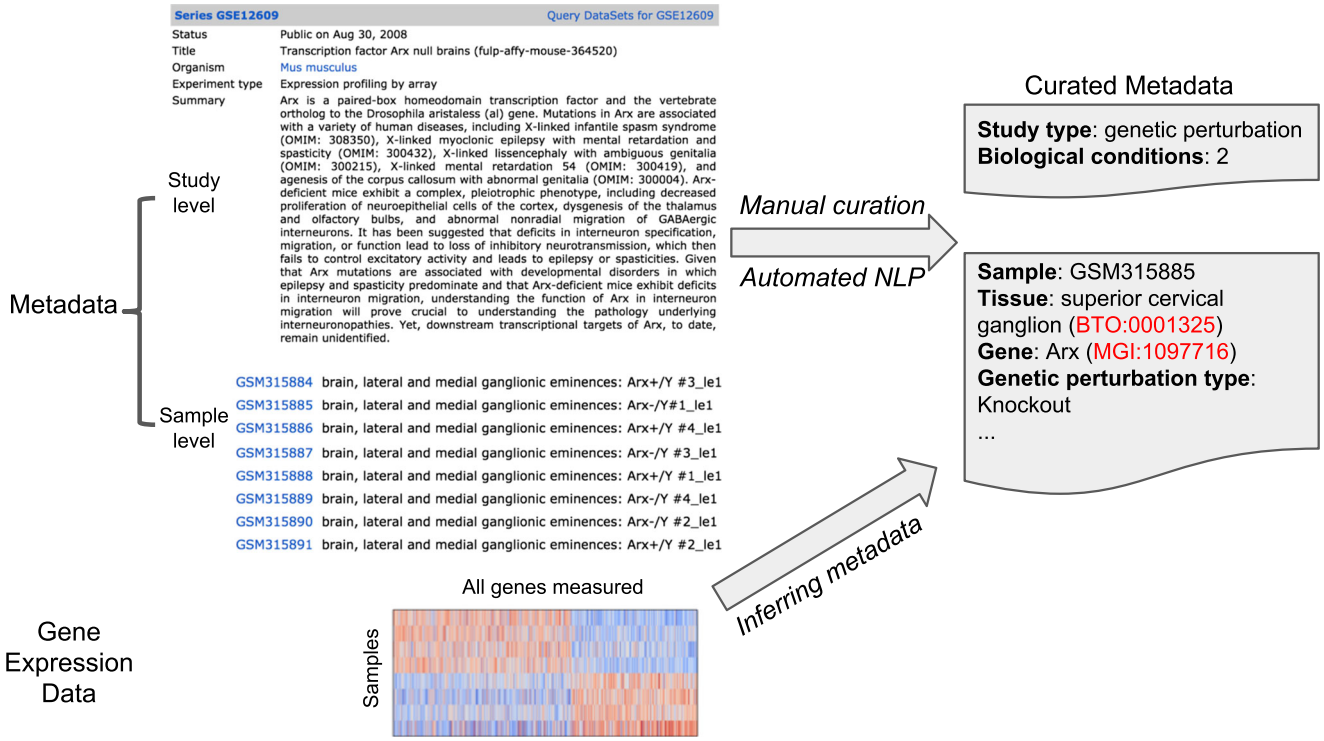

Fig. 2 Graphical summary of various curation approaches for further annotating GEO datasets. Metadata and the gene expression data from an example GEO study are shown on the left. Metadata are composed of semi-structured textual annotations supplied by the authors of the dataset at both study-level and sample-level to describe the experimental design of the study, and the characteristics of the samples. The goal of further annotating GEO datasets is to generate structured metadata for each study (top right) and samples (bottom right). Annotations are linked to relevant controlled vocabularies such as ontologies. Three approaches are visualized as arrows: manual curation and automated NLP, both attempt to identify and extract structured metadata from the textual descriptions. In addition, metadata can be inferred from the gene expression data using supervised machine learning approaches

to deal with errors made by manual curators is benchmarking. For instance, to validate the quality of the extracted signatures from STARGEO (Hadley et al. 2017), the authors showed that the DEGs from the meta-analysis of curated breast cancer datasets are comparable to signatures automatically generated from The Cancer Genome Atlas (TCGA) resource (The Cancer Genome Atlas Research N et al. 2013). Overall, manual curation efforts produce valuable resources to enable the systems pharmacology community.

Automated natural language processing

Applying natural language processing (NLP) techniques such as named-entity recognition (NER) and document classification to the textual descriptions of GEO studies is an attractive alternative for curating GEO metadata manually. NLP has been intensively applied to extract structured elements from the free-text of biomedical research publications over the past two decades (Huang and Lu 2016). Within this domain, NER is central. The goal of NER is to identify biological entities of interest, including genes, chemical/small-molecule/drug, disease, cell type, and tissue terms from free-text. Once key terms are identified, document classification models can be trained, using, for example, manually curated samples, to identify perturbation and control samples from GEO using labeled features from text identified by NER. Similarly, such document classification models can be trained to predict the themes of the datasets, including the specific drug treatment, disease model, or the genetic perturbation from the provided descriptions. We used the collection of the manually annotated CREEDS signatures metadata as a training set to train a document classifier for extracting the themes of the datasets from the entire GEO repository (Wang et al. 2016). Subsequent studies further improved NLP-based pipelines by enabling manual adjustments to the automatically curated gene expression datasets. For instance, GEOracle implements a machine learning (ML) classifier that identifies perturbation and control samples from GEO using textual features. It automatically tags samples as perturbation and controls to construct signatures. Importantly, it provides users with the ability to manually adjust the automated selection through a web interface (Djordjevic et al. 2017). Other related work

attempted to improve the general quality of the metadata associated with each sample and each GEO study. The leading effort is MetaSRA (Bernstein et al. 2017), a resource that normalized and improved the metadata from SRA. To achieve this, manual annotation of metadata applied to a small subset of SRA was carried out using ontologies for creating a training set. Then, by applying a computational model that implements a data structure called a Text Reasoning Graph, metadata labeling was automatically assigned to the remaining samples.

Inferring metadata from gene expression profiles

In addition to enriching and normalizing textual descriptions manually or automatically by examining the existing metadata, one can also leverage the information from the gene expression data itself to infer the metadata for curation. Given high-quality annotated gene expression profiles as a training set, ML models can be implemented to automatically identify the metadata from the gene expression profiles. For instance, various algorithms, including URSA (Lee et al. 2013), CIBERSORT (Newman et al. 2015), and xCell (Aran et al. 2017), were developed to predict cell types using gene expression data. Predicted cell types from such algorithms can be integrated with NER methods to corroborate the cell type terms recognized by NER to improve the accuracy of cell-type prediction algorithms directly from data. In the same way, other metadata elements can be predicted directly from the expression data. For example, the automated label extraction (ALE) (Giles et al. 2017) platform was used to impute the age, gender, and tissue type of samples from GEO using the expression data alone. Similarly to ALE, phenotype prediction of processed RNA-seq samples (Ellis et al. 2018) was implemented with ML methods trained using annotated samples from TCGA (The Cancer Genome Atlas Research N et al. 2013) and GTEx (Lonsdale et al. 2013). Another effort that utilized the Center for Expanded Data Annotation and Retrieval (CEDAR) framework (Panahiazar et al. 2017) tested the ability of a classifier to predict few basic common structured metadata elements such as cell type, organism, and platform from GEO samples.

Future perspectives

Further improving the curation of GEO datasets with deep and active learning

Current efforts in curating and annotating GEO datasets have exploited the information from both the textual descriptions and the gene expression profiles with manual crowdsourcing and automatic ML/NLP approaches. However, there is still room for further improving both the accuracy and the throughput of such curation tasks. Recent breakthroughs in NER were introduced by the application of deep learning (DL) for this task (Lample et al. 2016; Chiu and Nichols 2015). Due to the significant improved performance, such methods are currently considered the state-of-the-art. Deep neural network implementations of NER typically start with a word embedding layer that maps word tokens to low dimensional vectors that represent the meaning of the words learned from a large corpus using algorithms such as word2vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014). These word vectors are next connected to various long short-term memory (LSTM) or convolutional neural network (CNN) layers. Then, predictions can be made for each word token, suggesting whether the token is a start, a middle, or an end of a valid named-entity, or is an irrelevant token. The aforementioned state-of-the-art DL-based NER approaches have not been widely applied to biomedical data curation projects yet, perhaps with one exception (Habibi et al. 2017). In a recent study (Habibi et al. 2017), it was demonstrated that a deep neural network (DNN) model, specifically LSTM-Conditional Random Field (CRF) (Lample et al. 2016), outperforms domain-specific models with hand-crafted features in five biomedical NER tasks on 33 datasets. It would be promising to adopt the state-of-the-art deep NER algorithms, and train them on large biomedical corpora such as full-text articles from PubMed Central (PMC) to improve the accuracy of the mapped biological entities.

Another future direction to boost the quality and efficiency of the data curation task of GEO datasets is to develop a hybrid approach of manual and automated curation with active learning (AL). AL is a meta-algorithm for ML that learns to intelligently select examples (data points) for the underlying supervised ML algorithm to train and generalize more efficiently (Cohn et al. 1994). AL is particularly suitable for situations when unlabeled data is abundant and manual labeling is too expensive and time-consuming. AL algorithms attempt to overcome the lack of labeled data by asking human curators to aid with the labeling. The method strategically selects a subset of the data that needs labeling to maximally improve the model performance with minimal labeling requirement. This allows the ML algorithm to improve dynamically while reducing the effort necessary of the human curator (Krishnakumar 2007; Settles 2010). AL methods have been shown to achieve improved performance in similar crowdsourcing settings (Mozafari et al. 2014).

GEO dataset submission system with improved metadata standardization and validation

To prospectively improve the annotation quality of future datasets that will be deposited into GEO in the coming years, it would be a benefit to create a data and metadata submission system implemented with metadata standardization and

validation capabilities. It is feasible to implement web-based submission forms with metadata fields using various minimum information standards (Taylor et al. 2008) such as Minimum Information About a Microarray Experiment (MIAME) (Brazma et al. 2001). These fields can validate user input using external ontologies to ensure the accuracy of the deposited metadata. For instance, small molecule compounds used in a specific study can be validated by their chemical structure representation through UniChem (Chambers et al. 2013). Such mappings would enable cross-referencing to major public chemical databases to enrich the annotations by providing additional annotations, such as mechanism of actions, targets, disease associations, clinical phase status, and synonyms. It has been shown that such data submission systems, with deep metadata annotations that utilize established terminologies and ontologies, contribute to interoperability and reusability of the data (Stathias et al. 2018).

Toward making GEO datasets more FAIR

Recently, the findable, accessible, interoperable and reproducible (FAIR) guiding principles have been proposed to improve the groundwork needed to support the reuse of scientific data (Wilkinson et al. 2016). The ultimate goal of curating publicly available gene expression datasets is to make repositories such as GEO more FAIR. With the improved metadata annotations, GEO datasets will be more findable by both humans and machines through FAIR-compliant search engines such as the recently developed DataMed (Ohno-Machado et al. 2017; Chen et al. 2018) and Google DataSet Search (https://toolbox.google.com/datasetsearch). These search engines are powered by machine readable metadata that is hosted on dataset landing pages by the data repository using standards such as schema.org (Guha et al. 2016). Advances in web technologies also enable better interoperability between application programming interfaces (APIs). For instance, the BioThings APIs (Xin et al. 2018) can be cross-linked via JavaScript Object Notation for Linked Data (JSON-LD), a data format encoding semantically precise Linked Data, to enable automated knowledge extraction pipelines without having to specify the individual API endpoints and the returned data structures. The use of such technologies for building web services enables better interoperability, and can benefit the integration of GEO datasets with other resources and tools. For example, a researcher will be able to perform a drug-repurposing pipeline by simply specifying a disease of interest, to receive a ranked list of drugs as potential therapeutics through these web-services APIs. This pipeline will start by finding disease-related gene expression signatures, and then identify consensus DEGs through the API serving the annotated GEO datasets, which can then be applied as input for another API that serves drug repurposing queries such as those provided by the applications L1000CDS^2 (Duan et al. 2016), L1000FWD (Wang et al. 2018a), or clue.io (Subramanian et al. 2017) to retrieve a ranked list of drugs and compounds predicted to reverse the disease signature.

While the curation of metadata and the unified metadata models are important, optimal and uniform data processing pipelines, such as Recount2 (Collado-Torres et al. 2017), ARCHS4 (Lachmann et al. 2018), RNAseqDB (Wang et al. 2018b), and Toil Recompute (Vivian et al. 2017) are also vital for the reusability of the processed gene expression datasets. It is necessary to develop benchmarking strategies for processed datasets from different experimental and computational pipelines. For example, by comparing the consistency between transcription factor knockout and knockdown experiments with ChIP-seq studies that profiled the same transcription factors, we can evaluate the quality of RNA-seq alignment algorithms (Lachmann et al. 2018), calibrate the calling of genes from peaks for ChIP-seq studies, or benchmark methods for differential expression analysis (Clark et al. 2014).

Public gene expression data repositories such as GEO harbor enormous capacity for knowledge discovery. Outstanding progress has been achieved in developing methodologies and tools to facilitate the improved curation and reuse of those datasets in the past few years. However, there is still opportunity to develop better approaches to further advance the quality of GEO's metadata and data. With the FAIR guiding principles, the resultant improved curated public gene expression datasets will be integrated into an ecosystem of biomedical datasets and knowledge-bases for advancing biological discovery and for accelerating therapeutics development.

Funding information This work is supported by NIH grants U54-HL127624 (LINCS-DCIC), U24-CA224260 (IDG-KMC), and OT3-OD025467 (NIH Data Commons) to AM.

Compliance with ethical standards

Conflict of interest Zichen Wang declares that he has no conflict of interest. Alexander Lachmann declares that he has no conflict of interest. Avi Ma'ayan declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Aran D, Hu Z, Butte AJ (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 18(1):220

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41(D1):D991–D995

Bernstein MN, Doan A, Dewey CN (2017) MetaSRA: normalized human sample-specific metadata for the sequence read archive. Bioinformatics 33(18):2914–2923

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC et al (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29:365

Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, Hastings J, Bellis L, McGlinchey S, Overington JP (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform 5(1):3

Chen B, Butte A (2016) Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther 99(3):285–297

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14:128

Chen X, Gururaj AE, Ozyurt B, Liu R, Soysal E, Cohen T, Tiryaki F, Li Y, Zong N, Jiang M et al (2018) DataMed – an open source discovery index for finding biomedical datasets. J Am Med Inform Assoc 25(3):300–308

Cheng J, Yang L, Kumar V, Agarwal P (2014) Systematic evaluation of connectivity map for disease indications. Genome Med 6(12):95

Chiu JP, Nichols E (2015) Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint arXiv:151108308

Clark N, Hu K, Feldmann A, Kou Y, Chen E, Duan Q, Ma'ayan A (2014) The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics 15(1):79

Cohn D, Atlas L, Ladner R (1994) Improving generalization with active learning. Mach Learn 15(2):201–221

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT (2017) Reproducible RNA-seq analysis using recount2. Nat Biotechnol 35:319

Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23:1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Djordjevic D, Chen YX, Kwan SLS, Ling RWK, Qian G, Woo CYY, Ellis SJ, Ho JWK (2017) GEOracle: Mining perturbation experiments using free text metadata in Gene Expression Omnibus. bioRxiv

Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, Readhead B, Tritsch SR, Hodos R, Hafner M et al (2016) L1000CDS2: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl 2:16015

Dumas J, Gargano MA, Dancik GM (2016) shinyGEO: a web-based application for analyzing gene expression omnibus datasets. Bioinformatics 32(23):3679–3681

Ellis SE, Collado-Torres L, Jaffe A, Leek JT (2018) Improving the value of public RNA-seq expression data by phenotype prediction. Nucleic Acids Res 46(9):e54–e54

Giles CB, Brown CA, Ripperger M, Dennis Z, Roopnarinesingh X, Porter H, Perz A, Wren JD (2017) ALE: automated label extraction from GEO metadata. BMC Bioinformatics 18(14):509

Good BM, Su AI (2013) Crowdsourcing for bioinformatics. Bioinformatics 29(16):1925–1933

Guha RV, Brickley D, Macbeth S (2016) Schema. org: evolution of structured data on the web. Commun ACM 59(2):44–51

Gundersen GW, Jones MR, Rouillard AD, Kou Y, Monteiro CD, Feldmann AS, Hu KS, Ma'ayan A (2015) GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions. Bioinformatics. 31:3060–3062. https://doi.org/10.1093/bioinformatics/btv297

Gundersen GW, Jagodnik KM, Woodland H, Fernandez NF, Sani K, Dohlman AB, Ung PM-U, Monteiro CD, Schlessinger A, Ma'ayan A (2016) GEN3VA: aggregation and analysis of gene expression signatures from related studies. BMC Bioinformatics 17(1):461

Habibi M, Weber L, Neves M, Wiegandt DL, Leser U (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33(14):i37–i48

Hadley D, Pan J, El-Sayed O, Aljabban J, Aljabban I, Azad TD, Hadied MO, Raza S, Rayikanti BA, Chen B et al (2017) Precision annotation of digital samples in NCBI's gene expression omnibus. Sci Data 4:170125

Huang C-C, Lu Z (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform 17(1):132–144

Khare R, Good BM, Leaman R, Su AI, Lu Z (2015) Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform. 17:23–32. https://doi.org/10.1093/bib/bbv021

Kodama Y, Shumway M, Leinonen R (2012) The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 40(D1):D54–D56

Koeppen K, Stanton BA, Hampton TH (2017) ScanGEO: parallel mining of high-throughput gene expression data. Bioinformatics 33(21):3500–3501

Krishnakumar A (2007) Active learning literature survey. In.: Technical reports, University of California, Santa Cruz. 42

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44:W90–W97. https://doi.org/10.1093/nar/gkw377

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A (2018) Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 9(1):1366

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN et al (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313(5795):1929–1935

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural architectures for named entity recognition. arXiv preprint arXiv:160301360

Lee Y-s, Krishnan A, Zhu Q, Troyanskaya OG (2013) Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. Bioinformatics 29(23):3036–3044

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N et al (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45(6):580–585

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781

Mozafari B, Sarkar P, Franklin M, Jordan M, Madden S (2014) Scaling up crowd-sourcing to very large datasets: a case for active learning. Proceedings of the Very Large Data Bases Endowment 8(2):125–136

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B (2017) Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics 33(24):4033–4040

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA (2015) Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12:453

Ohno-Machado L, Sansone S-A, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Gururaj AE, Bell E et al

(2017) Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet 49:816

Panahiazar M, Dumontier M, Gevaert O (2017) Predicting biomedical metadata in CEDAR: a study of Gene Expression Omnibus (GEO). J Biomed Inform 72:132–139

Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 1532–1543

Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M et al (2013) ArrayExpress update—trends in database growth and links to data analysis tools. Nucleic Acids Res 41(D1):D987–D990

Settles B (2010) Active learning literature survey. University of Wisconsin, Madison 52(55–66):11

Shah N, Guo Y, Wendelsdorf KV, Lu Y, Sparks R, Tsang JS (2016) A crowdsourcing approach for reusing and meta-analyzing gene expression data. Nat Biotechnol advance online publication

Stathias V, Koleti A, Vidović D, Cooper DJ, Jagodnik KM, Terryn R, Forlin M, Chung C, Torre D, Ayad N et al (2018) Sustainable data and metadata management at the BD2K-LINCS Data Coordination and Integration Center. Sci Data 5:180117

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 102:15545–15550. https://doi.org/10.1073/pnas.0506580102

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK et al (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 171(6):1437–1452.e1417

Taylor CF, Field D, Sansone S-A, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz P-A, Bogue M, Booth T et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26:889

The Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, KRM S, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113–1120

Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME, Carmona-Sáez P (2018) ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty721

Torre D, Lachmann A, Ma'ayan A (2018) BioJupies: automated generation of interactive notebooks for RNA-Seq data analysis in the cloud. Cell Syst **7**(5):556–561.e553

Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, Pfeil J, Narkizian J, Deran AD, Musselman-Brown A (2017) Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol 35(4):314

Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, Jenkins SL, Feldmann AS, Hu KS, McDermott MG et al (2016) Extraction and analysis of signatures from the gene expression omnibus by the crowd. Nat Commun 7:12846

Wang Z, Lachmann A, Keenan AB, Ma'ayan A (2018a) L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. Bioinformatics 34:2150–2152. https://doi.org/10.1093/bioinformatics/bty060

Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, Zhang L, Minet T, Ochoa A, Gross BE, Iacobuzio-Donahue CA (2018b) Unifying cancer and normal RNA sequencing data from different sources. Sci Data 5:180061

Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39(suppl_2):W541–W545

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Xin J, Afrasiabi C, Lelong S, Adesara J, Tsueng G, Su AI, Wu C (2018) Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. BMC Bioinformatics 19(1):30

Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, Corney DC, Greene CS, Bongo LA, Kristensen VN et al (2015) Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nat Methods 12(3):211–214

Zinman GE, Naiman S, Kanfi Y, Cohen H, Bar-Joseph Z (2013) ExpressionBlast: mining large, unstructured expression databases. Nat Methods 10(10):925–926