Original Article

Healthc Inform Res. 2016 April;22(2):81-88.
http://dx.doi.org/10.4258/hir.2016.22.2.81
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

GEE: An Informatics Tool for Gene Expression Data Explore

Soo Youn Lee, PhD[1], Chan Hee Park, PhD[1], Jun Hee Yoon, MS[1], Sunmin Yun, MS[1], Ju Han Kim, MD, PhD[1,2]

[1]Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul, Korea; [2]Systems Biomedical Informatics–National Core Research Center (SBI-NCRC), Seoul National University College of Medicine, Seoul, Korea

Objectives: Major public high-throughput functional genomic data repositories, including the Gene Expression Omnibus (GEO) and ArrayExpress have rapidly expanded. As a result, a large number of diverse high-throughput functional genomic data retrieval systems have been developed. However, high-throughput functional genomic data retrieval remains challenging. **Methods:** We developed Gene Expression data Explore (GEE), the first powerful, flexible web and mobile search application for searching whole-genome epigenetic data and microarray data in public databases, such as GEO and ArrayExpress. **Results:** GEE provides an elaborate, convenient interface of query generation competences not available via various high-throughput functional genomic data retrieval systems, including GEO, ArrayExpress, and Atlas. In particular, GEE provides a suitable query generator using eVOC, the Experimental Factor Ontology (EFO), which is well represented with a variety of high-throughput functional genomic data experimental conditions. In addition, GEE provides an experimental design query constructor (EDQC), which provides elaborate retrieval filter conditions when the user designs real experiments. **Conclusions:** The web version of GEE is available at http://www.snubi.org/software/gee, and its app version is available from the Apple App Store.

Keywords: Microarray Analysis, Search Engine, RNA Sequence, Mobile Applications

I. Introduction

Major public repositories for microarray gene-expression data, including the Gene Expression Omnibus (GEO) [1] and ArrayExpress [2] have started accepting whole-genome epigenetic datasets created by RNA-seq and ChIP-seq technologies. These very large datasets present a challenge for efficient data retrieval system developers. Current tools for microarray data retrieval can be classified into a few categories (Table 1). The first category includes the Gene Expression Atlas [3], developed by the European Bioinformatics Institute (EBI), and GEO, developed by the National Center for Biotechnology Information (NCBI). The advantage of these tools is their ability to search for specific gene expression patterns under specific biological conditions. However, searching is only possible after all data have been pre-analyzed; therefore, these tools offer very limited search coverage of about 9% (Supplementary Figure 1). The second class, including GEO DataSets (http://www.ncbi.nlm.nih.gov/gds), developed by NCBI, and ArrayExpress, developed

Submitted: January 27, 2016
Revised: March 28, 2016
Accepted: April 8, 2016

Corresponding Author
Ju Han Kim, MD, PhD
Division of Biomedical Informatics, Systems Biomedical Informatics Research Center, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea. Tel: +82-2-740-8320, E-mail: juhan@snu.ac.kr

Table 1. Comparison of gene expression data retrieval tools

	Finding DEG & related experiments		Specific aims as collected data				Keyword search to rerated experiments			
	Atlas	GEO profiles	M^2DB	Microarray Retriever	GEO DataSets	Oncomine	GEOmetadb	GEO	ArrayExpress	GEE
With PubMed				●						
Using ontology	●				●			●	●	●
Local MySQL database							●			
R script		●					●			
Cancer specific						●				
Human curation annotation			●							
App tool										●
Conversion tool									●	●
Data size (experiment > 20,000)							●	●	●	●
Provided filter guide from human curated										●

DEG: Differentially Expressed Genes, GEO: Gene Expression Omnibus, M^2DB: microarray meta-analysis database, GEE: Gene Expression data Explore.

by EBI, which retrieve high-throughput functional genomic data based on the free-text metadata of experiments. These tools are divided into two types. The first type enables the user to input free-text keywords in a search system to search experiments related to specific biological conditions. The second type involves controlled vocabulary or ontology-based query systems. Given the vast diversity of biological conditions under which functional genomic datasets are created, queries are systematically guided by standardized terminologies for cell types, diseases, organism parts, etc., and their combinations may greatly improve performance and reduce the ambiguity of retrieval. These tools use a variety of controlled vocabulary or ontologies. Representative of GEO, ArrayExpress uses Medical Subject Headings (MeSH) [4] and Experimental Factor Ontology (EFO) [5]. MeSH allows the construction of terms that are easy to navigate and are useful to the search system because it consists of a hierarchy with 2013 trees (http://www.nlm.nih.gov/mesh/introduction.html). However, MeSH was created for indexing medical literature. Accordingly, MeSH does not include experiment-related terms such as “flow-sorted,” “Affymetrix Gene Chip ontology,” or “0 hour treatment.” Thus, MeSH is not suitable for controlled vocabulary when an experiment search is performed. EBI overcame the limitations of MeSH by creating the EFO to which they applied the ArrayExpress search system. EFO provides a systematic description of many experimental variables; however, this ontology does not have a suitable structure for term navigation when an experiment search is conducted, because each category only has a depth of 2–4 (http://purl.bioontology.org/ontology/EFO) experimental conditions. For example, EBI tools (Atlas, ArrayExpress) provide only four filter conditions.

We aimed to overcome the limitations of previous retrieval tools by developing Gene Expression data Explore (GEE), which uses eVOC and EFO to overcome the limitation of previous controlled-vocabulary-based search systems. eVOC is a controlled vocabulary for unifying gene expression data [6]. eVOC and EFO are representative ontologies for the description of high-throughput functional genomics data. The eVOC and EFO ontologies include essential terms for the efficient retrieval of high-throughput functional genomic datasets. eVOC is an ontology that associates labeled target cDNAs for microarray experiments, or cDNA libraries and their associated transcripts with controlled terms in a set of hierarchical vocabularies and consists of 4 orthogonal controlled vocabularies, including anatomical system, cell type, pathology, and developmental stage. eVOC has a well-defined classification structure for term navigation; however, it only includes 2,260 terms. On the other hand, EFO has about 7,000 terms, but a poor classification structure. Thus, we merged eVOC and EFO by retaining the structure of eVOC to reflect the search system of GEE. GEE provides a specific advanced search system called the experiment design query constructor (EDQC), which provides five categories and 13 filter conditions, including “antibody”, and includes 1,516 terms. Previous search systems provide only web applications; however, GEE provides the first mobile application. Thus, users are able to perform an experiment search regardless of time and place. The GEE website is http://www.snubi.org/software/gee, and the GEE app can be downloaded from the Apple App Store.

II. Methods

1. Data Collection and Management

GEE covers all transcriptomic platforms, including RNA-seq and ChIP-seq, as well as different types of microarray technologies. GEE archives 31,245 high-throughput functional genomic datasets with 710,038 samples extracted from GEO and ArrayExpress, i.e., 978 (911 samples; 2,408 platform types) and 30,267 (709,127 samples; 8,127 platform types) experiments from ArrayExpress and GEO, respectively. Additionally, 2,720 highly curated GEO DataSets (GDS) are included (Supplementary Figure 1A). GEE has the largest data scope (standard date: 2012.01.01).

2. eVOC Ontology-Based Annotation of Datasets

We downloaded the OBO file format of the eVOC ontology and the owl format of the EFO ontology from the bioportal website (http://bioportal.bioontology.org) [7]. We extracted terms, synonyms, and associated information. Next, stop words were removed and executed. Stemming and normalization of eVOC terms, EFO terms, and all of the synonyms were performed by application of the Porter stemming algorithm. Then, we merged 2,264 eVOC terms and 5,081 terms from EFO using the bioportal mapping list (Supplementary Table 1), after which we mapped EFO to eVOC because the eVOC structure is well defined. We annotated all the high-throughput functional genomic datasets with ontology terms, which were processed using text mining. We annotated the datasets by using a MySQL full-text search method with four filter conditions. A different query method was applied to each category of terms (Supplementary Figure 2). Our work showed that the use of a full-text search method with 4 filter conditions and different query methods using eVOC and EFO is the best approach among several methods we tried for mapping (Supplementary Figure 3).

3. User Interface Feature

1) SEARCH option in web version of GEE

The GEE web interface contains two sections for generating search queries (Figure 1A). Section I offers an ontology-based keyword search function, which enables a user to select multiple ontology terms from an ontology term tree. Section II contains an EDQC, which enables a user to generate a query when designing experiments for high-throughput functional genomic data. The EDQC values were human curated. The EDQC consists of 3 classes, 5 parts, and 14 subparts. The sample condition class is composed of the sample extraction type, used anti-body, gene modification, organism, sex, and population. This class includes information associated with samples. The analysis class is composed of data normalization and data quality controls. This class includes information associated with data analysis. The platform class is composed of protocol manufacture, kinds of arrays, platform technology, platform support, platform coating, and platform hardware. This class included platform information, such as experiments using ChIP-seq or RNA-seq. All selected components are performed independently. Users can generate multiplex combination queries using all kinds of classes and terms. Furthermore, these queries provide correct results according to user-defined logic, and redundant results are removed. More information can be found in Figure 2B.

2) SEARCH option in app version of GEE

The GEE app enables the retrieval of high-throughput functional genomic datasets using eVOC ontology terms. The app provides a user friendly methodology. Initially, the user would have to download and install the GEE app from the Apple App Store. Next, the user uses one click to select the search term from among the eVOC ontology terms. When the user clicks the selected term, the GEE app provides a high-throughput functional genomic list of datasets from which a dataset name could be selected to obtain more information about the specific genomics. As a result, the GEE app displays more detailed information when the user selects the dataset. The app interface is shown in more detail in Figure 1B. The GEE app is currently under review in the Apple App Store.

3) GEE web and app interface construction of GEE

The GEE web version was created by using Hypertext Markup Language 5 (HTML5), cascading style sheets 3 (CSS3)

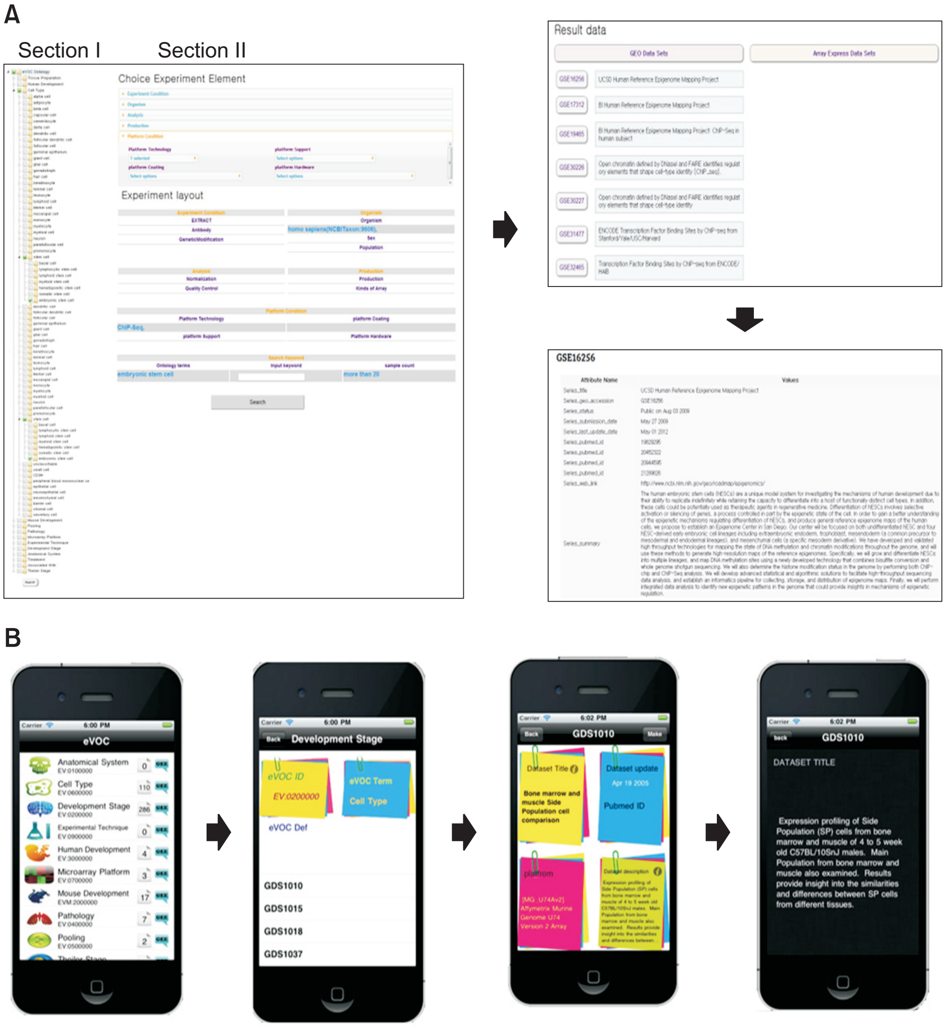


Figure 1. Example search (search parameters are representative in example page). (A) is the web version of Gene Expression data Explore (GEE) and (B) is the app version of GEE.

and jQuery. The web version of GEE is available on the web at http://matrix.snubi.org:8080/GEE_index.jsp, and the app version of GEE is available from the Apple App Store.

4. Comparing Search Performance of Other Search Tools to GEE

We used four keywords ("breast cancer," "stem cell," "lung cancer," and "myocardial infarction"), which are the most frequently submitted terms in the PubMed query log [8]. We used our proposed tool to extract 500 GDS datasets randomly from among the GDS datasets. In addition, the same 4 keywords were used to manually extract related datasets from the 500 GDS datasets, after which the intersection dataset was extracted from the human curated datasets. We used GEO, ArrayExpress, and GEE to search this dataset using the 4 keywords obtained from the 500 datasets, after which we again obtained the intersection dataset with human curated datasets. Finally, we calculated the search performance using precision, by recalling the F-measure method (Supplementary Table 2).

5. SOFT and MAGE-TAB Conversion

GEE provides MAGE-TAB [9] and SOFT conversion. This enhances the efficiency of the retrieval process by allowing the use of filter conditions and specific queries. We extracted all attributes from MAGE-TAB. Specification ver. 1.1 has been made available together with SOFT guidelines (http://www.ncbi.nlm.nih.gov/geo/info/soft2.html), and the relation rules were created manually (Supplementary Table 3). The results provided on the GEE website are based on the results obtained by using the conversion rules. Overall, our scheme maps the GEO series (GSE) attributes to investigation description format (IDF), sample and data relationship format (SDRF), and raw and processed data files are mapped to GEO samples (GSM), and array design format (ADF) is mapped to GEO platform (GPL). A detailed description is included as Supplementary Table 4.

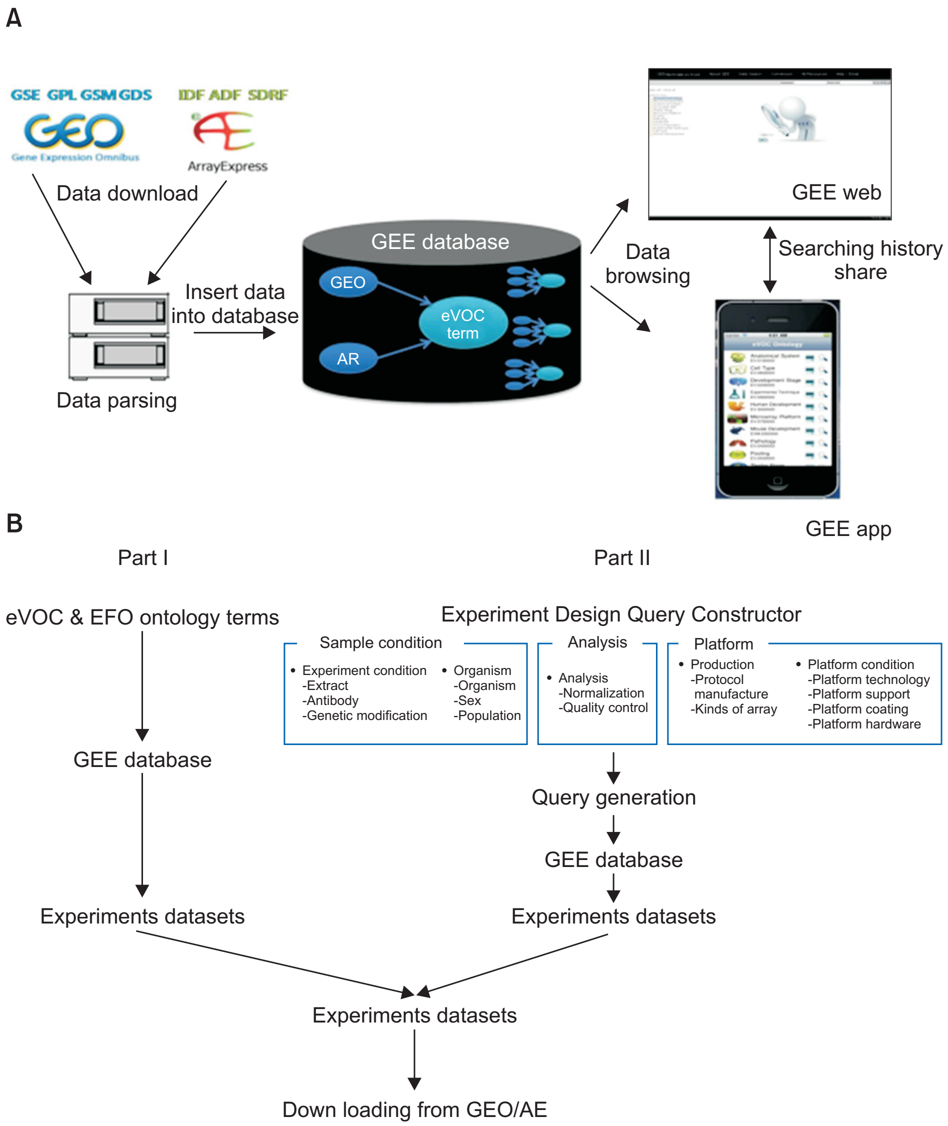


Figure 2. System (A) and search schema (B) of Gene Expression data Explore (GEE).

III. Results

1. Combined Effect of Using Ontology and EDQC

GEE uses an EDQC and an ontology-based query system to search high-throughput functional genomic data by using a pre-annotated data table. We calculated the mapping coverage in GEE according to the method described in Supplementary Table 2. We tested three conditions, including using only ontology (Supplementary Figure 4A), using only EDQC (Supplementary Figure 4B), and using ontology with EDQC (Supplementary Figure 4C and D). The results were calculated by each data format. Most data formats had more than 95% mapping coverage. Unfortunately, ADF was obtained when using ontology, with GSE in EDQC being the exception. However, all the conditions improved when ontology was combined with EDQC (Supplementary Figure 4C and D). This enabled us to conclude that a combination of ontology and EDQC is effective for searching high-throughput functional genomic data.

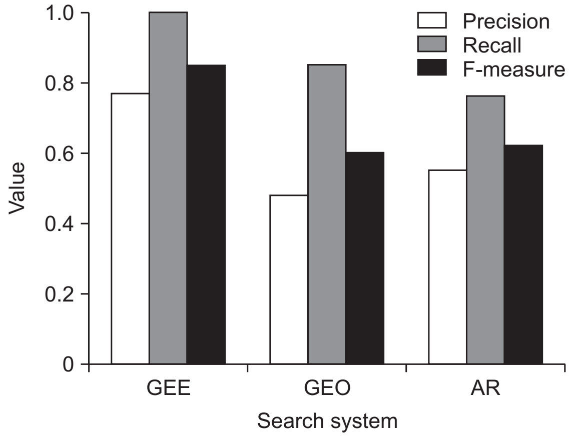
2. Comparing the Search Performance of Previous High-Throughput Functional Genomic Data Searching System and GEE

We calculated the precision, recall, and F-measure to compare the searching performance of GEO, ArrayExpress, and GEE. We randomly extracted 500 GDS datasets and used 4 keywords for sample queries. The keywords that were extracted were those most frequently submitted in the PubMed query log [8]. Details of the method are described in the Method section and Supplementary Table 2. The most important aspect of a search system is to search for the correct data in user defined queries and to ensure an abundance of results. These conditions can be satisfied by ensuring harmony between precision and the recall result. GEE produced a high F-measure score. The testing results show that GEE achieved the best score for all performance measurements compared to other tools (Figure 3). In particular, the recall result was perfect. This result shows that GEE searches the correct data in any biological query. Moreover, the F-measure is the most effective for testing, which means that GEE searches more accurately than GEO and ArrayExpress.

3. Example Query Using GEE

As shown in Figure 1A, the result of using the web version of GEE is that only the elements that are the most critical are shown and provide a good overview of the datasets in GEO or ArrayExpress with a hyperlink to each dataset. A search example using the web version of GEE is shown in Figure 1A. We demonstrated the performance of the web version of GEE by retrieving a particular sample query using the EDQC. The first sample query was executed by using “ChIP-seq,” “embryonic stem cell,” and “homo sapiens,” containing more than 20 samples in the keywords from one dataset. “ChIP-seq” and “homo sapiens” contained more than 20 keyword samples from the EDQC in Section II and “embryonic stem cell” from the ontology term tree in Section I (Figure 1A). GEO, ArrayExpress, and GEE each produced 6 datasets as results (GSE36114, GSE32970 ... etc., 984 samples, 5 platform), 0 result datasets, and 18 experiment datasets as results (GSE30226, GSE30227 ... etc., 536 samples, 0 platforms) in response to the query.

The second sample query was based on the use of “anaplastic astrocytoma,” and “eye neoplasm” as keywords. Both keywords are from the ontology tree in Section I (Figure 1A). GEO, ArrayExpress, and GEE each produced 0 datasets and 63 datasets (GSE7330 ... etc., 921 samples, 0 platform). However, ArrayExpress did not support sample count filtering, and the sample data link was broken. GEE provides a description of each dataset and its download link. Therefore, the user can download data when the dataset permission is opened.



	GEE	GEO	AR
Precision	0.77	0.48	0.55
Recall	1.00	0.85	0.76
F-measure	0.85	0.60	0.62

Figure 3. Search performance comparison of previous high-throughput functional genomic data searching systems and Gene Expression data Explorer (GEE). GEO: Gene Expression Omnibus, AR: ArrayExpress.

The app version of GEE consists of two parts. Part I provides easy retrieval of high-throughput functional genomic data using an eVOC ontology tree with just one touch. When the user selects a term, the GEE app provides a list of related datasets, and term information appears when the user selects the dataset. Part II displays a summary of the information and a detailed description of the selected datasets. The resulting sample is presented in Figure 1B.

IV. Discussion

GEE has four main benefits, which will be discussed in detail in this section. First GEE enables specific queries to be made using eVOC and EFO vocabularies and the EDQC system. Moreover, it is possible to retrieve suitable high-throughput functional genomic datasets consisting of free text from diverse biological conditions. Biologists often design and execute new experiments because they are unable to easily find suitable experiment datasets. Therefore, many duplicate datasets are generated and saved in GEO and ArrayExpress. We expect all of these problems to be resolved through the EDQC of GEE because the EDQC will be able to provide suitable queries when users design experiments.

Second, GEE had the largest dataset scope, currently based on August 1, 2012. GEE archives 31,245 high-throughput functional genomic datasets (710,038 samples) from GEO and ArrayExpress. This huge data scope, including whole-genome epigenetic datasets, provides a wealth of resulting datasets upon retrieval. Beside GEE, there are several tools with which to retrieve datasets, such as the Microarray Retriever [10], GEOmetadb [11], M2DB [12], and Oncomine [13]. However, they do not provide whole-genome epigenetic datasets, such as RNA-seq and ChIP-seq. Therefore, there is no need for the user to simultaneously access both repositories when GEE is used. In addition, GEE provides a searching result for each platform together with sample information. This function provides elaborate information of high-throughput functional genomic datasets.

Third, GEE provides two optimized user systems: a web application of GEE (http://www.snubi.org/software/gee) and a mobile app of GEE. Every high-throughput functional genomic data retrieval system simply consists of web-based applications. However, GEE also provides a mobile app, which is convenient, rapid, and portable for the user. The GEE app is the first mobile application to search high-throughput functional genomic datasets. Assuming that users have smartphones, they will be provided with suitable high-throughput functional genomic datasets after just one click of the mobile app of GEE; thus, the mobile app allows the user flexibility.

Fourth, GEE is available with specific filter conditions and retrieval using MAGE-TAB and SOFT, the high-throughput functional genomic dataset formats of AR and GEO, respectively, with attributes such as platform technology, and a sample treatment protocol based on the MAGE-TAB and SOFT converter. These attributes allow the user to specify conditions for specific queries, and these queries contribute to the retrieval of accurate datasets in large high-throughput functional genomic datasets. As mentioned previously, GEE is the only tool that enables the retrieval of whole-genome epigenetic datasets and has a mobile app. GEE is also the only application that offers an elaborate searching system that uses the EDQC and ontology trees. We expect GEE (http://www.snubi.org/software/gee) to promote the evolution of retrieving high-throughput functional genomic datasets.

In the coming years, the number of whole-genome epigenetic and microarray datasets is expected to increase explosively. Therefore, the problems associated with the retrieval of large high-throughput functional genomic datasets are becoming increasingly important. GEE can play a key role in solving this problem.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2012-0000994) and the ICT R&D program of MSIP/IITP (No. B0101-15-247, Development of open ICT healing platform using personal health data).

Supplementary Materials

Supplementary materials can be found via http://dx.doi.org/10.4258/hir.2016.22.2.81. Figure 1. Comparison of related tools data coverage (anchor date: 2012.01.01). Figure 2. Different query method by each keyword category. Figure 3. Comparison of annotation methods in GDS datasets. (A) Results from using 4 filter conditions, including some terms containing numbers (“8 cells”), addresses (“http:mged.sourceforge.net/ontologies”), symbol (“C2 Mouse”), and chemicals ("3-(p-Hydroxyphenyl)alanine”). (B) Results from the testing query performance when filter conditions were added one by one. Figure 4. A combination effect of using ontology and EDQC. (A) Mapping coverage when ontology is used (eVOC with EFO). (B) Mapping coverage when the EDQC is used. (C, D) Mapping coverage when ontology is used with EDQC (dark blue indicates unmapped data count, and sky blue indicates mapped data count). Table 1. Number of terms in mapping processing EFO and eVOC. Table 2. Measurement for search efficiency. Table 3. Comparison of filter conditions in various search systems. Table 4. Conversion rule for MAGE-TAB and SOFT.

References

1. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 2009;37(Database issue):D885-90.
2. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, et al. ArrayExpress update: an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 2011;39(Database issue):D1002-4.
3. Cheng WC, Tsai ML, Chang CW, Huang CL, Chen CR, Shu WY, et al. Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. BMC Bioinformatics 2010;11:421.
4. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, et al. Gene Expression Atlas update: a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res 2012;40(Database issue):D1077-81.
5. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. J Am Med Inform Assoc 2001;8(4):317-23.
6. Malone J, Holloway E, Adamusiak T, Kapushesky M,

Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 2010;26(8):1112-8.

7. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, et al. eVOC: a controlled vocabulary for unifying gene expression data. Genome Res 2003;13(6A): 1222-30.
8. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009; 37(Web Server issue):W170-3.
9. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 2006;7:489.
10. Lu Z, Wilbur WJ, McEntyre JR, Iskhakov A, Szilagyi L. Finding query suggestions for PubMed. AMIA Annu Symp Proc 2009;2009:396-400.
11. Ivliev AE, 't Hoen PA, Villerius MP, den Dunnen JT, Brandt BW. Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. Nucleic Acids Res 2008;36(Web Server issue): W327-31.
12. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics 2008;24(23):2798-800.
13. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007;9(2): 166-80.

Supplementary Materials

Healthc Inform Res. 2016 April;22(2):81-88.
http://dx.doi.org/10.4258/hir.2016.22.2.81
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

A

	GEO	Array Express	GEE	GEO metadb	GEO dataset	Atlas	GEO profiles
Experiment	29,134	27,796	31,245	27,401	2,720	3,476	2,720
Sample	143,954	157,339	141,825	136,046	14,725	99,305	14,725
Platform	9,938	2,960	10,535	9,799	401	408	401

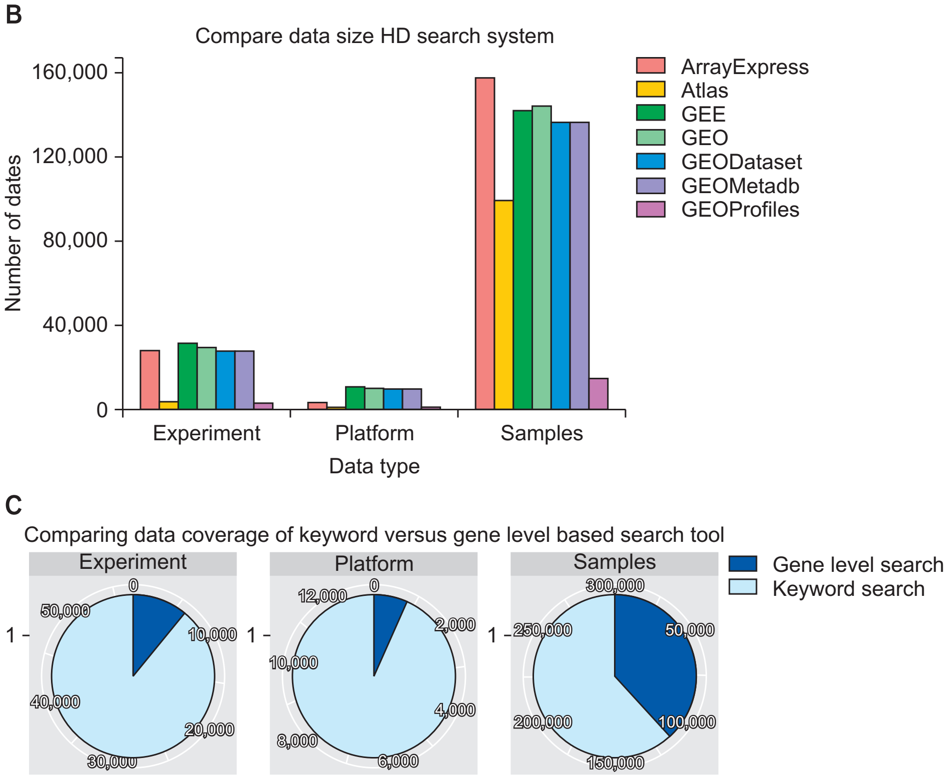
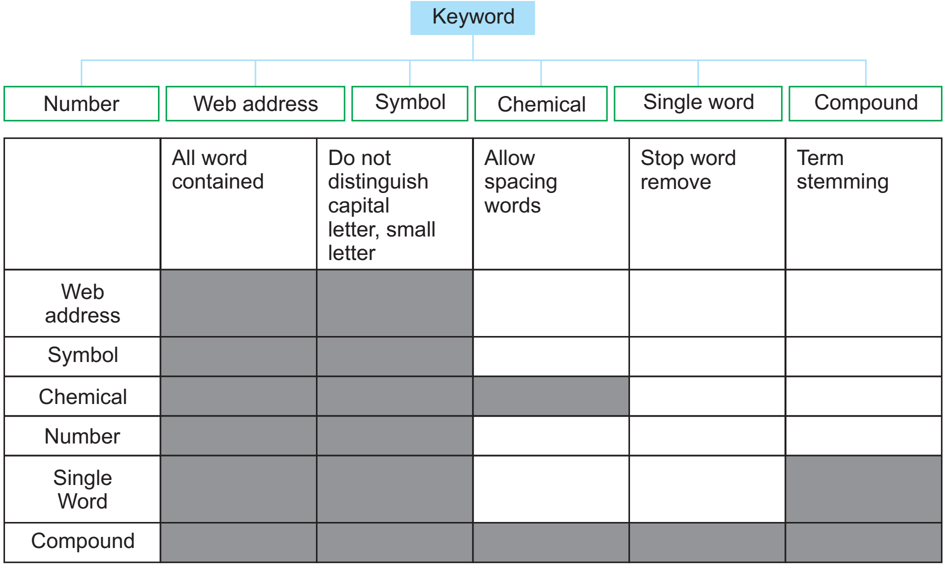


Figure 1. Comparison of related tools data coverage (anchor date: 2012.01.01).

Table 1. Number of terms in mapping processing EFO and eVOC

Attributes	No. of terms
eVOC + EFO (total _ID)	6,838
Intersection eVOC and EFO terms	908
eVOC terms	2264
EFO terms	5081
No. of syms	20,950



	All word contained	Do not distinguish capital letter, small letter	Allow spacing words	Stop word remove	Term stemming
Web address	▨	▨			
Symbol	▨	▨			
Chemical	▨	▨	▨		
Number	▨	▨			
Single Word	▨	▨			▨
Compound	▨	▨	▨	▨	▨

Figure 2. Different query method by each keyword category.

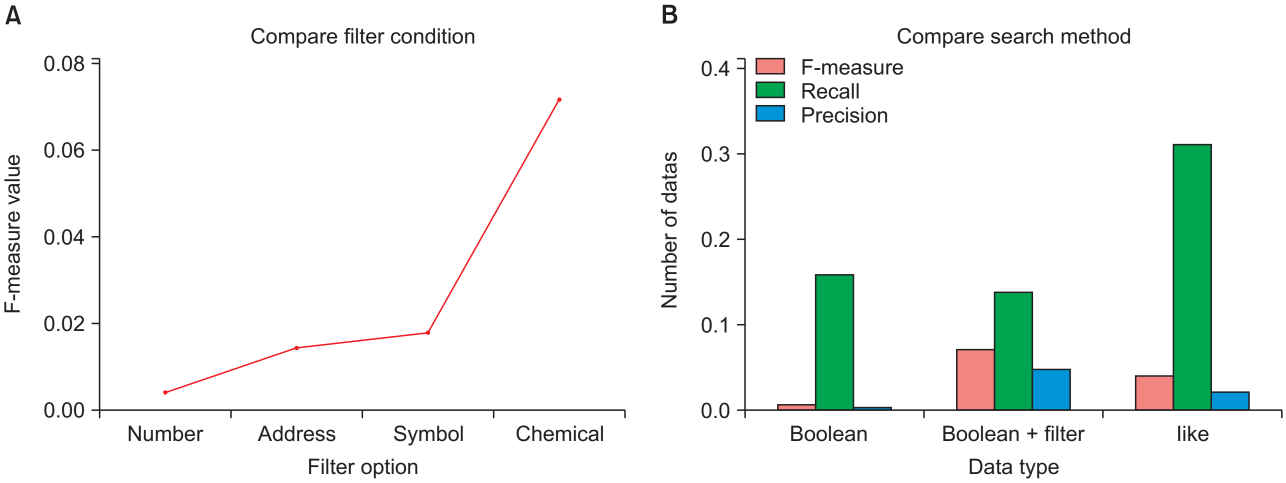


Figure 3. Comparison of annotation methods in GDS datasets. (A) Results from using 4 filter conditions, including some terms containing numbers (“8 cells”), addresses (“http:mged.sourceforge.net/ontologies”), symbol (“C2 Mouse”), and chemicals ("3-(p-Hydroxyphenyl)alanine"). (B) Results from the testing query performance when filter conditions were added one by one.

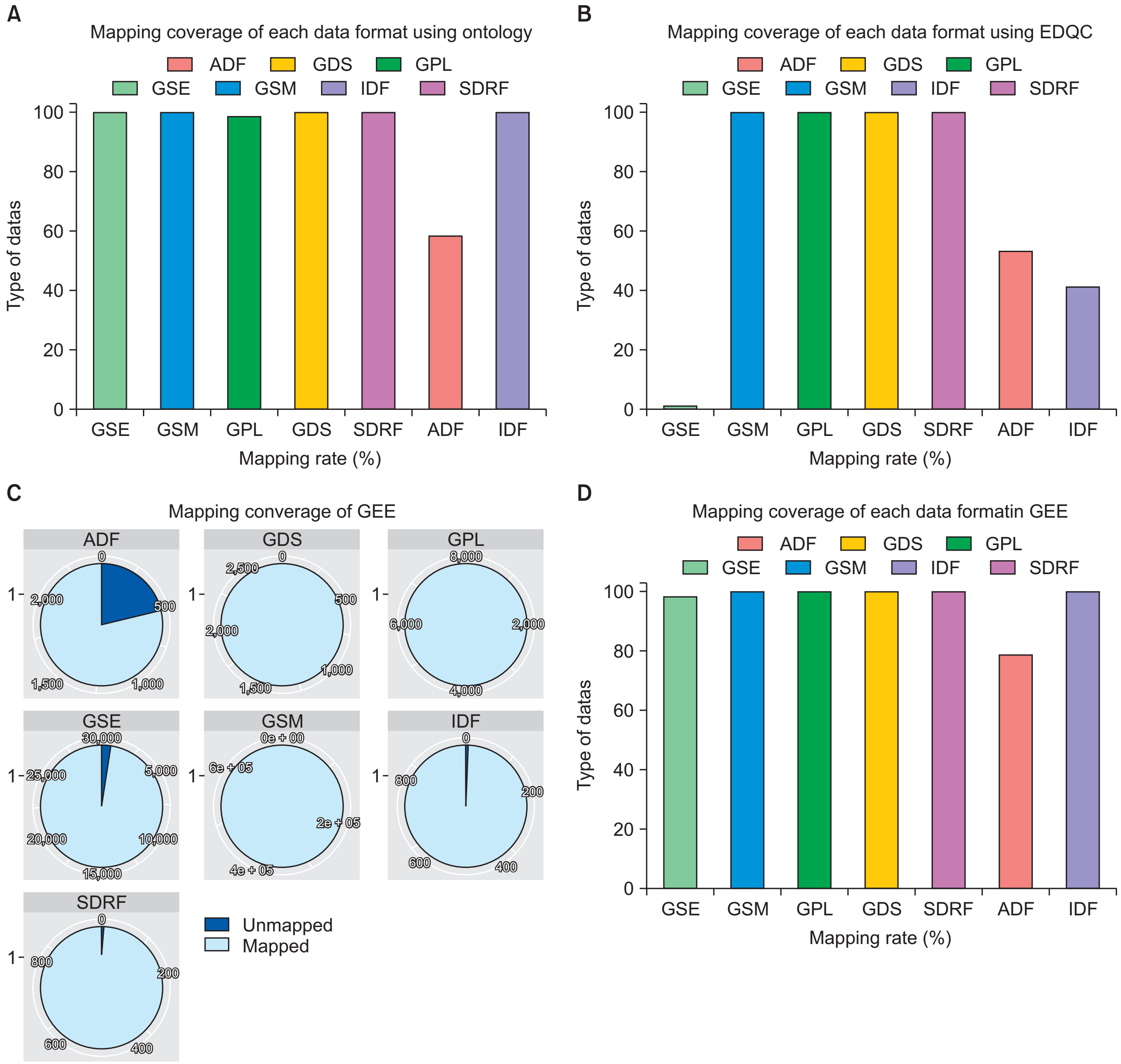


Figure 4. A combination effect of using ontology and EDQC. (A) Mapping coverage when ontology is used (eVOC with EFO). (B) Mapping coverage when the EDQC is used. (C, D) Mapping coverage when ontology is used with EDQC (dark blue indicates unmapped data count, and sky blue indicates mapped data count). EDQC: experiment design query constructor, EFO: experimental factor ontology. GEE: Gene Expression data Explore, GEO: Gene Expression Omnibus, GSE: GEO series, GSM: GEO samples, GPL: GEO platform, GDS: GEO datasets, SDRF: sample and data relationship format, ADF: array design format, IDF: investigation description format.

Table 2. Measurement for search efficiency

Precision = $\frac{\lvert\{\mathit{relevant\ document}\}\cap\{\mathit{retrieved\ document}\}\rvert}{\lvert\{\mathit{retrieved\ document}\}}$
Recall = $\frac{\lvert\{\mathit{relevant\ document}\}\cap\{\mathit{retrieved\ document}\}\rvert}{\lvert\{\mathit{total\ retrieved\ document}\}}$
F-measure = $2 \times \frac{\mathit{precision} \times \mathit{recall}}{\mathit{recision} + \mathit{recall}}$
Mapping coverage = $2 \times \frac{\mathit{count}\ (\mathit{distinct\ mapped\ dataset})}{\mathit{count}\ (\mathit{number\ of\ dataset})} \times 100$

Table 3. Comparison of filter conditions in various search systems

GEO	AR	GEE
Study sample number	Array (2,972)	Experiment condition
Author	All assays by molecule (5)	-Extract (13)
Dates	All technologies (3)	-Genetic modification (387)
Organism	Species (1,493)	-Antibody (54)
Study type		Organism
Supplementary files		-Organism (842)
Subset variable type		-Sex (2)
-Age		-Population (54)
-Agent		Analysis
-Cell line		-Normalization (8)
-Cell type		-Quality control (11)
-Development stage		Production
Attribute name		-Protocol manufacture (30)
Dataset type		Platform condition
Description		-Platform technology (26)
Entry type		-Platform coating (24)
Filter		-Platform support (16)
GEO accession		-Platform hardware (49)
Mesh term		
Number of platform probes		
Number of samples		
Organism		
Platform technology		
Project		
Sample type		
Sample variable type		
Submitter institute		
Subset description		
Subset variable type		
Tag length		
Title		
Sample source		
Related platform		
Relate series		
Reporter identifier		

Values in parentheses are number of kinds of filter conditions.

GEO is free-text-based search system, and AR and GEE are system that provides search term for search.

GEO: Gene Expression Omnibus, AR: ArrayExpress, GEE: Gene Expression data Explore.

Table 4. Conversion rule for MAGE-TAB and SOFT

MAGE-TAB attributes	GEO attributes
MAGE-TAB version	
Investigation title	Series_title
	dataset_title
Experimental design	Series_overall_design
	Series_summary
	dataset_description
Experimental design term source REF	
Experimental design term accession number	
Experimental factor name	Series_variable_description_1
Experimental factor type	Series_type
	dataset_type
Experimental factor term source REF	Series_variable_1
Experimental factor term accession number	Series_variable_sample_list_1
Person last name	Series_contributor, Series_contact_name
Person first name	
Person mid initials	
Person email	Series_contact_email
Person phone	Series_contact_phone
Person fax	Series_contact_fax
Person address	Series_contact_laboratory
	Series_contact_department
	Series_contact_institute
	Series_contact_address
	Series_contact_city
	Series_contact_state
	Series_contact_zip/postal_code
	Series_contact_country
Person affiliation	
Person roles	
Person roles term source REF	
Person roles term accession number	
Quality control type	
Quality control term source REF	
Quality control term accession number	
Replicate type	Series_repeats_1
Replicate term source REF	Series_repeats_sample_list_1
Replicate term accession number	
Normalization type	
Normalization term source REF	
Normalization term accession number	
Date of experiment	
Public release date	Series_submission_date
	Series_last_update_date
	dataset_update_date
PubMed ID	Series_pubmed_id
Publication DOI	
Publication author list	
Publication title	
Publication status	
Publication status term source REF	
Publication status term accession number	
Experiment description	Series_summary

Table 4. Continued 1

Protocol name	Platform_manufacture_protocol
	Sample_treatment_protocol_ch[n]
	Sample_growth_protocol_ch[n]
	Sample_label_protocol_ch[n]
	Sample_hyb_protocol
	Sample_scan_protocol
	Sample_extract_protocol_ch[n]
Protocol type	
Protocol term source REF	
Protocol term accession number	
Protocol description	
Protocol parameters	
Protocol hardware	
Protocol software	
Protocol contact	
SDRF file	Series_sample_id, Series_contact_address
Term source name	Sample_geo_accession
Term source file	
Term source version	
Comment []	Series_geo_accession
	Series_web_link
	Series_supplementary_file
	Series_platform_id
	Series_platform_taxid
	Series_sample_taxid
	Series_relation
	Series_citation
	Series_contact_web_link
	Database_name
	Database_institute
	Database_web_link
	Database_email
	Database_ref
	dataset_reference_series
	dataset_order
Array design name	Platform_title
	Platform_description
	dataset_platform
Version	Annotation_date
Provider	Platform_manufacturer
	Platform_web_link
	Platform_catalog_number
	Platform_contributor
	Platform_contact_name
	Platform_contact_name
	Platform_catalog_number
	Platform_contact_phone
	Platform_contact_institute
	Platform_contact_city
	Platform_contact_state
	Platform_contact_zip/postal_code
	Platform_contact_country
	Platform_contact_web_link
	Platform_catalog_number
	Platform_contributor
	Platform_contact_fax
	Platform_contact_department
	Platform_contact_address
Printing protocol	Platform_technology
	dataset_platform_technology_type

Table 4. Continued 2

Technology type		Platform_distribution
Technology type term source REF		
Technology type term accession number		
Surface type		Platform_coating,Platform_support
Surface type term source REF		
Surface type term accession number		
Substrate type		subset_description subset_sample_id subset_type
Substrate type term source REF		
Substrate type term accession number		
Sequence polymer type		
Sequence polymer type term source REF		
Sequence polymer type term accession number		
Term source name		
Term source file		
Term source version		
Comment []		Platform_support
		Platform_organism
		Platform_geo_accession
		Platform_pubmed_id
		Platform_status
		Platform_submission_date
		Platform_last_update_date
		Platform_taxid
		Platform_relation
		Platform_supplementary_file
		Platform_citation
		dataset_platform dataset_platform_organism dataset_feature_count Annotation_platform Annotation_platform_title Annotation_platform_organism
Source name	Characteristics	Sample_source_name_ch(1,2)
	Provider	Sample_biomaterial_provider_ch(1,2)
		Sample_contact_name
		Sample_contact_email
		Sample_contact_phone
		Sample_contact_fax
		Sample_contact_laboratory
		Sample_contact_department
		Sample_contact_institute
		Sample_contact_address
		Sample_contact_city
		Sample_contact_state
		Sample_contact_zip/postal_code
		Sample_contact_country
	Material type	
	Description	
	Comment	
Sample name	Characteristic	Sample_characteristics_ch(1,2)
	Material type	Sample_type
		dataset_sample_type
	Description	Sample_title
		Sample_description
	Comment	Sample_organism_ch(1,2)
		dataset_sample_organism
		Sample_status
		Sample_submission_date
		dataset_channel_count
		dataset_sample_count
		Sample_last_update_date

Table 4. Continued 3

Extract name	Characteristics, Material type, Description, Comment	Sample_molecule_ch(1,2)
Labeled extract name	Characteristics, Material type, Description, Label, Comment	Sample_label_ch(1,2)
Hybridization name	Array data file, Derived array data file, Array data matrix file, Derived array data matrix file, Array design file / REF, Technology type, Comment	Sample_channel_count
Assay name	Technology type, Array data file, Derived array data file, Array data matrix file, Derived array data matrix file, Array design file / REF, Comment	
Scan name	Array data file, Derived array data file, Array data matrix file, Derived array data matrix file, Comment	Sample_supplementary_file
Normalization name	Derived array data file, Derived array data matrix file, Comment	Sample_data_processing dataset_value_type
Array data file	Comment	Sample_platform_id
Derived array data file	Comment	
Array data matrix file	Comment	
Derived array data matrix file	Comment	
Image file	Comment	
Array design file / REF	Term source REF, Comment	Sample_platform_id Sample_taxid_ch(1,2)
Protocol REF	Term source REF, Parameter, Performer, Date, Comment	Sample_treatment_protocol_ch(1,2) Sample_extract_protocol_ch(1,2) Sample_label_protocol_ch1 Sample_hyb_protocol Sample_scan_protocol
Comment []		Sample_series_id Sample_data_row_count

Color represents the kind of format. Pink is for GSE attributes, gray is for GSM, and blue is for GPL, dark blue is for GDS, light green is for IDF, yellow is for SDRF, and purple is for ADF.

GEO: Gene Expression Omnibus, GSE: GEO series, GSM: GEO samples, GPL: GEO platform, GDS: GEO datasets, IDF: investigation description format, SDRF: sample and data relationship format, ADF: array design format.