Published in final edited form as:

Methods Mol Biol. 2006 ; 338: 175–190.

Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO)*

Tanya Barrett and Ron Edgar

Summary

The Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) has emerged as the leading fully public repository for gene expression data. This chapter describes how to use Web-based interfaces, applications, and graphics to effectively explore, visualize, and interpret the hundreds of microarray studies and millions of gene expression patterns stored in GEO. Data can be examined from both experiment-centric and gene-centric perspectives using user-friendly tools that do not require specialized expertise in microarray analysis or time-consuming download of massive data sets. The GEO database is publicly accessible through the World Wide Web at <http://www.ncbi.nlm.nih.gov/geo>.

Keywords

Microarray; gene expression; database; data mining

1. Introduction

Microarray technology is one of the most important experimental developments in molecular biology in recent years. Microarrays have enabled researchers to conduct large-scale quantitative assessments of gene expression, defining the transcriptome of a multitude of cellular types and states.

The National Center for Biotechnology Information (NCBI) launched the Gene Expression Omnibus (GEO) database in 2000 to support the public use and dissemination of gene expression data generated by high-throughput methodologies (*1,2*). The database is populated by material supplied by the scientific community. Most researchers submit to GEO in accordance with grant or journal requirements stipulating that microarray data be made available through a public repository, in compliance with long-established standards of scientific reporting that allow others to judge or reproduce the results. Consequently, most of the data presented in GEO has been analyzed and published. GEO is not intended or suitable for initial analysis of newly acquired data, which is typically the role taken by laboratory information management systems (LIMS).

The GEO database stores molecular abundance data generated by a wide variety of high-throughput measuring techniques. These include microarray-based experiments that measure gene expression or detect genomic gains and losses (comparative genomic hybridization), as well as genomic tiling arrays that are used to detect transcribed regions or single-nucleotide polymorphisms, or to identify protein-binding genomic regions in conjunction with chromatin immunoprecipitation (ChIP-chip technology). Some non-array-based high-throughput data types are also accepted by GEO, including serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS), serial analysis of ribosomal sequence tags

*This chapter is an official contribution of the National Institutes of Health; not subject to copyright in the United States.

(SARST), and some peptide profiling techniques such as tandem mass spectrometry (MS/MS). The data analysis features discussed here are generally applicable to all these technology types, but for the purposes of this chapter the focus is on microarray-generated gene expression data, which currently constitute about 95% of the data in GEO.

At the time of writing, GEO holds over 50,000 submissions, representing approximately half a billion individual molecular abundance measurements, for over 100 organisms, submitted by over 1000 laboratories. These data explore a huge breadth of biological phenomena, for example, mouse models of diabetes, flower development in plants, anthrax sporulation, aging in fruit flies, effect of cigarette smoke on bronchial cells, kidney transplant rejection, toxicological effects of antimalarial drugs, and many others. When one is working with a vast compendium of data, it is important to be able to effectively query the data, focusing on those that are relevant to a specific area of interest. This chapter describes intuitive interfaces and tools that help researchers effectively explore, visualize, and interpret the submitted data. These tools do not require specialized knowledge of microarray analysis methods, nor do they require time-consuming download of large data sets.

2. Organization of the Database

To readily interpret the data in GEO, it helps to have a general understanding of the database structure and content. Researchers provide their data in four sections: Platform, Sample, Series (which receive persistent GPLxxx, GSMxxx, and GSExxx accession numbers, respectively), and raw data. A Platform defines the array template and contains sequence identity tracking information for each feature on the array. A Sample record contains the measured hybridization data, along with a description of the biological source and treatment protocols. A Series record ties together experimentally related Samples. Accompanying raw data (e.g., Affymetrix original probe data or cDNA array.tif images) may be optionally supplied.

The hardware and software packages that generate and process microarray data produce a wide assortment of data styles and formats—Platform and Sample tables can take on many different structures and contain multiple and varying types of ancillary and supporting information. Furthermore, microarray-based technologies and processing strategies are continually evolving. The GEO database was designed with these considerations in mind and has a flexible and open architecture that can accommodate variety. However, data provided in different styles and formats are not readily interpretable or analyzable, even by the experienced user. To address this issue, an upper level of organization is applied. Despite the diversity, a common core of relevant data is supplied to GEO:

- Sequence identity tracking information for each feature on the array.
- Normalized hybridization measurements.
- A description of the biological source used in each hybridization.

These data are extracted from the submitter-supplied records and reassembled by GEO staff into an upper level unit called a GEO DataSet (assigned a persistent GDSxxx accession). A DataSet represents a collection of similarly processed, experimentally related hybridizations. Samples within a DataSet are further organized according to experimental variable subgroups, for example, they are categorized by age, disease state, and so on. A DataSet can be rendered to generate two separate representations of the data:

1. An *experiment-centered* perspective that encapsulates the whole study. This information is presented as a *DataSet record*. DataSet records comprise a synopsis of the experiment, a breakdown of the experimental variables, access to several data display and analysis tools, and download options (Fig. 1).

2. A *gene-centered* perspective that provides quantitative gene expression measurements for individual genes across the DataSet. This information is presented as a *GEO Profile*. GEO Profiles comprise gene identity annotation, the DataSet title, and a chart depicting value and rank measurements for that gene across all Samples in that DataSet (Fig. 2).

DataSets represent a standardized format for the data in GEO. All the data analysis and mining tools described in this chapter are based on DataSets.

3. Retrieving and Analyzing GEO Data

GEO DataSets may be browsed at

http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi. Original submitter-supplied records may be browsed at http://www.ncbi.nlm.nih.gov/projects/geo/info/print_stats.cgi. A specific record may be accessed directly using its assigned GEO accession number. Additionally, all data (DataSets, original submitter-supplied data, and raw data) are freely available for bulk download via FTP at <ftp://ftp.ncbi.nlm.nih.gov/pub/geo/data/>.

The utility of vast quantities of data such as these is increased greatly if there are effective methods to query the database, allowing reduction to data that are relevant to a specific area of interest. Several different query approaches have been developed, including standard text-based searches, nucleotide sequence-based searches, queries based on characteristics of the expression patterns themselves, or combinations of these parameters.

3.1. Searches Using NCBI's Entrez Search Engine

Most biological scientists are familiar with Entrez, using it routinely to search NCBI databases like PubMed and GenBank (3,4). It has a straightforward interface by which users can simply type in key words and terms to locate relevant material, as well as capabilities to perform complex searches across multiple databases. The Entrez system currently comprises 25 interlinked databases, two of which store GEO data:

1. *Entrez GEO DataSets* contains DataSet definitions. Researchers can search for DataSets using text key word terms. Retrievals display the DataSet title, a synoptic description, the organism, and the experimental variables, as well as links to the complete GDS record (Fig. 1), parent Platform, and reference Series records. A user can quickly scan through the retrievals and identify DataSets that look relevant to the area of study. Entrez GEO DataSets is searchable using the "DataSets" query box on the GEO home page, or at <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds>.
2. *Entrez GEO Profiles* contains individual, normalized, DataSet-specific gene expression profiles. Researchers can query for specific genes by name, symbol, accession number, and clone identifier, or genes of interest based on characteristics of the expression profiles. Retrievals display the mapped gene name (determined by the sequence reference provided on the array), the DataSet title, and a thumbnail chart of the gene expression profile generated from the normalized values for an individual gene reporter in each Sample of the DataSet. Clicking on the thumbnail image will enlarge the chart to display the full profile details and the Sample subset partitions that reflect experimental design (Fig. 2). Entrez GEO Profiles is searchable using the "Gene profiles" query box on the GEO home page, or at <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo>.

It is usually sufficient to simply enter text key words to retrieve data of interest. For example, searching GEO Profiles with the term "Klk13" will retrieve all profiles that match that gene name. However, information is indexed into several categories (Table 1). These indices can

be used to refine and restrict Entrez queries. To perform such a search, specify the search terms, their fields, and the Boolean operations to perform on the term using the following syntax:

term [field] OPERATOR term [field]

where term(s) are the search terms, the field(s) are the search fields and qualifiers, and the OPERATOR(s) are the Boolean operators (uppercase AND, OR, NOT). The Preview/Limits link on the Entrez tool bar assists greatly in construction of complex queries. Alternatively, complex search statements can be written and executed directly in the search boxes. The indices (available on the Preview/Limits page) may be used to browse and/or select the terms by which the data are described.

3.1.1. Identifying Experiments of Interest—In many cases, a researcher will begin by looking for DataSets that are pertinent to the area of study. For example, to locate experiments that investigate spermatogenesis or testis development in mice using Affymetrix GeneChip technology, search Entrez GEO DataSets with “(spermatogenesis OR testis development) AND Affymetrix AND mouse.”

3.1.2. Identifying Gene Expression Profiles of Interest—Once the researcher has located relevant DataSet(s), he or she can use the DataSet accession number (GDSxxx) to restrict searches to that experiment. For example, to view the profiles of all heat shock genes in DataSet GDS181, he or she could query with “GDS181 AND heat shock.”

If a DataSet accession is not specified, then the search will be performed across all GEO data. For example, to view profiles of kallikrein family genes in any DataSet that investigates progesterone, enter “kallikrein AND progesterone[GDS Text].”

Often the gene name will contain words that can be found in the DataSet description, and vice versa. In these cases, the only way to specifically retrieve data would be to restrict the search using the appropriate qualifier. For example, a researcher trying to locate profiles for the glycine receptor gene would have to restrict the search to “glycine receptor[Gene Description]”; otherwise, he or she would also retrieve all profiles from GDS967, a DataSet investigating a “Glycine receptor beta subunit mutant model for hyperekplexia.”

Several fields are available for refining a search to help identify interesting or significant profiles based on the characteristics of the expression pattern. For example, the expression measurements of each Sample in a DataSet are rank ordered. It is possible to refine searches to identify genes that fall within a specified abundance bracket. To view profiles that fall into the top 5% abundance rank bracket in at least one Sample in DataSet GDS182, search with “GDS182 AND 96:100[Max Value Rank].” Alternatively, to search for profiles in which the median value level across the DataSet is high (approx 12–14 for this DataSet), query with “GDS182 AND 12:14[Median value in profile].”

Typically, researchers seek genes that vary their expression depending on experimental factors. As described in **Subheading 2.**, GEO DataSets are partitioned into subsets that reflect experimental design. Profiles are flagged if they display a significant effect in relation to subsets, that is, if the expression values or ranks pass a threshold of statistical difference between any non-single experimental variable subset and another. These flags assist in the identification of candidate genes as follows:

- Users can restrict their searches to find any profile that exhibits an effect within specific DataSets. For example, to view profiles showing interesting value subset effects in either DataSet GDS186 or GDS187, query with “(GDS186 OR GDS187)

AND “value subset effect” [Flag Type].” A convenient way to run this search is from the DataSet record page (*see* subset effect box in Fig. 1).

- Users can search across the whole GEO for genes that show an effect with respect to a particular experimental variable type. For example, to search for any gene that shows an effect with respect to gender, query with “gender [Flag Information].”
- Standard GEO Profile retrievals are default ordered according to subset effect flags, bringing potentially significant and interesting profiles to the fore.

The Entrez search system is a powerful tool that interlinks many diverse data domains. Regular users of NCBI resources are well advised to familiarize themselves with the advanced mining features available through Entrez (*see* **Note 1**).

3.2. Tools Available Within Entrez GEO Profiles Results Page

After identification of profile(s) of interest, there are several features on the profile records that link to various types of related profiles, assisting in identification of more genes of interest, or to related information in other NCBI Entrez databases:

Profile neighbors connects genes that show a similar or reversed profile shape within a DataSet, as calculated by Pearson correlation coefficients. Profile neighbors may suggest that those genes have a coordinated transcriptional response, possibly inferring some common function or regulatory elements (*see* **Note 2**).

Sequence neighbors retrieves profiles related by nucleotide sequence similarity across all DataSets, as determined by BLAST (5). Retrievals are shown in decreasing order of similarity to the selected sequence and can provide valuable insights into the possible function of the original sequence if it has not yet been characterized, or they may be useful in identifying related gene family members or for cross-species comparisons.

Homolog neighbors retrieves profiles of genes belonging to the same Homolo-Gene group. HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.

Links menu allows users to easily traverse from the GEO databases to associated records in other Entrez data domains including GenBank, PubMed, Gene, UniGene, OMIM, HomoloGene, Taxonomy, SAGEMap, and MapViewer.

3.3. Query Mean Group A vs B

The “Query mean group A vs B” feature is available on DataSet records (Fig. 1). This tool enables researchers to identify genes that have specific profile characteristics with regard to experimental factors within a DataSet. The user assigns one or multiple subsets to group A and other subset(s) to group B. He or she can then specify that he wants to retrieve all profiles in which the mean value or rank is, for example, 4+-fold higher in group A compared with group B and is directed to Entrez GEO profiles that match those criteria.

For example, for DataSet GDS279, if a researcher wants to locate genes that display a 3+-fold increase in expression between mice that were fed high-fat diet compared with mice that were fed a low-fat diet, he or she would check boxes, as indicated in Fig. 1. Hitting the “Query A vs B” button would retrieve profiles that meet these specifications.

3.4. Cluster Heat Maps

One of the most powerful methods to mine and visualize high-dimensional data is through cluster analyses. Details on the mathematical basis of these clustering algorithms are not within

the scope of this chapter, but simply speaking, cluster analyses attempt to detect natural groups in data using a combination of distance metrics and linkages. GEO provides nine classic varieties of precomputed unsupervised hierarchical clusters, as well as user-defined K-means and K-median clustering (Fig. 1). Columns (Samples), and independently, the rows (genes) are rearranged to place rows with similar response patterns near each other and columns with similar response patterns near each other. Cluster results are graphically represented as “heat maps,” whereby high through low expression levels are presented as a two-color spectrum that allows the user to easily identify groups of interesting genes through visual pattern recognition. Each distinct colored “island” in the heat map represents a coordinated transcriptional response, based on the assumption that genes having similar expression profiles across a set of conditions are likely to be involved in the same biological processes. Such biologically relevant clusters can lead to the formulation of testable predictions and can infer functional roles for previously uncharacterized genes.

The GEO cluster heat map images are interactive; using a moveable box, users can select a region, or regions, of interest. This region can be enlarged and the raw data downloaded, plotted as line charts, or linked out to the corresponding profiles in Entrez GEO Profiles (Fig. 3; see Note 3).

3.5. GEO BLAST

The GEO BLAST feature is available on the GEO home page and from NCBI's BLAST page (<http://www.ncbi.nlm.nih.gov/BLAST/>). This tool allows retrieval of expression profiles on the basis of BLAST nucleotide sequence similarity. Researchers paste in a nucleotide sequence, or specify a sequence accession number, and a BLAST query is performed against all GenBank identifiers represented on microarray Platforms or SAGE libraries in GEO. The output resembles conventional BLAST output with each alignment receiving a quality score; each retrieval has an expression “E” icon that links directly to corresponding GEO Profiles. This interface is helpful in identifying sequence homologs, e.g., related gene family members or for cross-species comparisons, or for providing valuable insight into possible roles of the original sequence if it has not yet been functionally characterized.

4. Conclusions

The GEO database archives large volumes of gene expression data generated by the scientific community. Several different approaches to mining GEO data are outlined in this chapter; each of these methods assists biologists to drill down through inherently noisy expression data to genes that are relevant, or behave in a way that is relevant, to their particular area of study.

Making such a large collection of data accessible and analyzable using common interfaces adds a valuable investigative dimension not attained when considering isolated experiments. Through analyzing multiple, independently generated DataSets that examine similar phenomena, it is possible to substantiate interesting gene expression trends that may have been overlooked, or are borderline, in one experiment alone. Researchers can look to see what the preponderance of evidence indicates about the behavior of a gene, or group of genes (7,8). Users can mine GEO for evidence that corroborates laboratory findings, or they may look to GEO for candidate genes worthy of further study in the laboratory. Having sequence information together with expression information can help in the functional annotation and characterization of unknown genes, or in finding novel roles for characterized genes. These data are also valuable to genome-wide studies, allowing biologists to review global gene expression in various cell types and states, to compare with orthologs in other species, and to search for repeated patterns of coregulated groups of transcripts that assist formulation of hypotheses on functional networks and pathways (9–11).

Additionally, integration of GEO data into NCBI's Entrez search engine greatly expands the utility of the data. Entrez is a powerful tool that enables disparate data in multiple databases to be richly interconnected. This can lead to inference of previously unidentified relationships between diverse data types, facilitating novel hypothesis generation, or assisting in the interpretation of available information. Such opportunities for discovery will only increase as the database continues to grow.

The GEO database is under continuous development, so the examples and data presentation strategies described in this chapter may become outdated over time. To keep informed of the latest GEO developments, subscribe to the GEO mailing list at geo@ncbi.nlm.nih.gov.

5. Notes

1. Advanced mining tips using Entrez searches.
 - a. Use "History" in the Entrez tool bar to see your previous queries. Each search is assigned a number and is stored for up to 8 h. Previous queries can be combined to form a new search query.
 - b. Use the "Display" pull-down menu to find related data in other Entrez resources. For example, let us say that your GEO Profiles search has narrowed down to a list of 100 candidate genes. You next want to check the Gene Expression Nervous System Atlas (GENSAT) database for complementary expression evidence for those genes. (GENSAT is another NCBI resource that contains expression mapping information for genes in the mouse brain at various stages of development.) Instead of checking each of the 100 genes individually in GENSAT, you would simply select "GENSAT Links" from the Display menu in GEO Profiles, and you are immediately directed to GENSAT data that corresponds to your 100 candidate genes.
 - c. GEO mining tools, together with the Entrez features described in **steps a** and **b** above, can be combined to form very powerful searches. For example, consider DataSets GDS214 and GDS563—these are independent experiments performed on different arrays that compare normal muscle tissue with muscle tissue from patients affected with Duchenne muscular dystrophy. A user could perform the following set of maneuvers to identify genes that are upregulated in Duchenne patients, in both DataSets.
 - i. Use cluster analysis in GDS214 to visually select clusters of genes that are highly expressed in Duchenne Samples compared with control.
 - ii. Use the "Get Profiles" button to export these genes to Entrez GEO Profiles.
 - iii. Select "Gene Links" from the "Display" pull-down menu—this retrieves a list of corresponding curated genes from NCBI's "Gene" database. From the History tab, you can see that this search is assigned #1.
 - iv. Repeat the above three steps for GDS563—from the History tab, you can see that this search is assigned #2.
 - v. Combine these two searches by querying Entrez Gene with "#1 AND #2." This retrieves a list of common genes that are found to be upregulated in Duchenne patients in two separate

DataSets. The fact that these genes appear to be similarly regulated in both DataSets lends confidence to the results. This also demonstrates a way to effectively perform cross-platform analyses.

- d. The “MyNCBI” feature allows users to save searches and retrieve them in a later session, or monitor how a prior saved search is modified in the context of the current, updated database content. To use the many features of MyNCBI, the user must first establish a login name in that system.
 - e. Profile neighbor links are subject to cutoff limit. Thus, if this limit is reached, bear in mind that there are probably more genes in the DataSet that demonstrate similar behavior. In this case, you might consider utilizing cluster analyses or the “Query mean group A vs B” tool, which are not subject to such limitations.
2. It is important to realize that different cluster methods will generate different results. An underlying assumption of clustering is that genes with similar expression patterns are more likely to have similar biological function. Clustering does not provide proof of this relationship, but it does provide suggestions for data interpretation.
 3. The gene expression value bars are plotted on the left y-axis. Note that this scale slides to fit the values of a particular profile. This sliding scale allows subtle differences in values to be more clearly visualized. The ranks are plotted on the right y-axis and are always scaled from 0 to 100%.
 4. Binned rank information is provided as complementary indication of the relative abundance of a gene compared with all other genes on that array. A rank profile that follows the trend of the corresponding value profile provides additional assurance that the data are properly normalized. Keep in mind that cross-gene rank assessments are made with the assumption that all probes are detecting their target with the same efficiency, which may not always be true.
 5. The Samples within any comparable DataSet are assumed to have been processed similarly. You can verify that Sample values are well distributed and normalized with respect to each other (and thus comparable) by viewing the “value distribution” chart that is provided on each DataSet record under the “analysis” button. This presents a box and whisker plot for each Sample within the DataSet, allowing easy visualization of the value median, spread, and overall range.
 6. The “Sort” button on GEO Profile charts lets users resort the Samples in the DataSet according to a particular experimental variable. This can assist in clearer visualization of an expression trend in experiments with multiple variables.

Acknowledgements

The authors unreservedly acknowledge the expertise and dedication of the GEO curation and development team—Carlos Evangelista, Pierre Ledoux, Dmitry Rudnev, Alexandra Soboleva, Tugba Suzek, Dennis Troup, and Steve Wilhite.

References

1. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res* 2005;33(Database issue):D562–566. [PubMed: 15608262]
2. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207–210. [PubMed: 11752295]
3. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005;33(Database issue):D39–45. [PubMed: 15608222]

4. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 1996;266:141–162. [PubMed: 8743683]
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. [PubMed: 2231712]
6. Recinos A 3rd, Carr BK, Bartos DB, et al. Liver gene expression associated with diet and lesion development in atherosclerosis-prone mice: induction of components of alternative complement pathway. *Physiol. Genomics* 2004;19:131–142.
7. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. *Science* 2003;300:1749–1751. [PubMed: 12805549]
8. Zerbini LF, Wang Y, Czibere A, et al. NF-kappa B-mediated repression of growth arrest- and DNA-damage-inducible proteins 45alpha and gamma is essential for cancer cell survival. *Proc Natl Acad Sci USA* 2004;101:13618–13623. [PubMed: 15353598]
9. Rodwell GE, Sonu R, Zahn JM, et al. A transcriptional profile of aging in the human kidney. *PLoS Biol* 2004;2:e427. [PubMed: 15562319]
10. Scott MS, Perkins T, Bunnell S, Pepin F, Thomas DY, Hallett MT. Identifying regulatory subnetworks for a set of genes. *Mol Cell Proteomics*. 2005 Feb 18;Epub
11. Haverty PM, Frith MC, Weng Z. CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Res* 2004;32(Web Server issue):W213–216. [PubMed: 15215383]

The screenshot displays the NCBI GEO Data Set record for GDS279. The interface includes a top navigation bar with links to NCBI Handbook, Chapter 1, NAR 2002 Paper, FAQ, MIAME, and Email GEO. The main content area is divided into sections: GDS Summary, Subset and Sample Info, and a list of 12 samples.

GDS Summary:

- Accession: GDS279 [View Expression Profiles \(Entrez-GEO\)](#)
- Title: Diet induced changes in liver
- Summary: Examination of gene expression profiles from liver of C57BL/6 wild type mice and LDL receptor deficient mice fed on either a low fat diet or a high fat, western-style diet for 12 weeks.
- Organism: Mus musculus
- Platform: GPL81: Affymetrix GeneChip Murine Genome U74 Version 2 Set MG-U74A
- Experiment type: single channel nucleotide
- No. of probes: 12488
- Value type: count
- Series: GSE363
- Series published: April 05 2003
- Last GDS update: December 08 2003

Subset and Sample Info:

Sample selection: ☒ check all ☐ uncheck all ☐ toggle

Data: ☒ download ☐ analysis

4 assigned subsets:

Sample	Type	Description
☒ (6)	strain	LDL receptor deficient mutant
☒ (6)	strain	C57BL/6 wild type
☒ (6)	protocol	low fat diet
☒ (6)	protocol	high fat diet

☒ GDS279 only ☒ ranks ☒ values ☐ subset effects

Query mean group A vs B by values:

A	3+ fold	higher	B
☐	☐	☐	☐
☐	☐	☐	☐
☐	☐	☐	☐
☒	☐	☐	☐
☐	☐	☐	☐

☒ Query A vs. B ☐

12 samples, order: none

☒ GSM5403: C57_JF_1 src: liver	☒ GSM5404: C57_JF_2 src: liver	☒ GSM5405: C57_JF_3 src: liver	☒ GSM5406: C57_JF_4 src: liver
☒ GSM5410: LDL_JF_2 src: liver	☒ GSM5411: LDL_JF_3 src: liver	☒ GSM5406: C57_JF_1 src: liver	☒ GSM5407: C57_JF_2 src: liver
☒ GSM5408: C57_JF_3 src: liver	☒ GSM5412: LDL_JF_1 src: liver	☒ GSM5413: LDL_JF_2 src: liver	☒ GSM5414: LDL_JF_3 src: liver

Annotations:

- Download Data Set value matrix or annotation matrix
- Access to all clusters and value distribution chart
- Quick link to hierarchical clusters
- Query mean group A vs B tool
- Experimental subgroup variables
- Individual Samples that constitute the Data Set
- Subset effect search tool

Fig. 1. GEO Data Set record. A screen shot of a typical Data Set record, GDS279, which investigates the effect of a high-fat diet on liver tissue in wild-type and LDL receptor-deficient mice (6). The locations of the main Data Set features and tools are indicated.

Fig. 2.

Screenshot of GEO Profiles retrievals and expanded profile chart. **(A)** Screen shot of GEO Profiles retrievals for GDS279, a DataSet that investigates the effect of a high-fat diet on liver tissue in wild-type and LDL receptor-deficient mice (6). Each retrieval represents an individual gene on the array. The locations of various features are indicated. **(B)** Expanded chart for the *Cidec* gene. The chart bars represent relative gene expression levels in each Sample across the DataSet (see Note 4). The squares represent rank information; ranks are calculated by rank-ordering expression values for all the genes that are present on the array for each Sample and then placing them into percentile bins. Thus, ranks give an indication of the level of expression of that gene compared with all other genes on the array (or they can be an indication of when data are not properly normalized; see Notes 5 and 6). The bars at the foot of the chart reflect experimental design (see Note 7). From the chart, it is clear that expression of *Cidec* is upregulated in both strains of mice fed a high-fat diet.

Fig. 3. DataSet cluster analysis. Section of DataSet GDS279 uncentered correlation UPGMA hierarchical cluster analysis. Each column represents an individual Sample, or hybridization; each row represents a gene, identified by a GenBank accession number. The light color indicates high expression and the darker color low expression. The dashed box can be moved and resized to select regions of interest, the data for which may be downloaded, or exported to Entrez GEO Profiles.

Table 1Entrez Qualifier Fields^a

Field name	Field description
GEO DataSets	
Author	Authors associated with the experiment
Experiment Type	The experiment type, e.g., cDNA, genomic, protein, SAGE
GDS Text	DataSet description text
GEO Accession	The GEO accession number (GPLxxx, GSMxxx, GSExxx, GDSxxx)
GEO Description/Title Text	Text provided in the description/title of original records
Number of Samples	The number of Samples in the DataSet [*]
Number of Platform Probes	The number of Platform reporters in the DataSet [*]
Organism	The organism from which the reporters on the array were derived/designed
Reporter Identifier	The identifier for the array reporter (GenBank accession, gene name, and so on)
Sample Source	The source biological material of the Sample
Sample Title	Sample Title
Submitter Institute	Submitter institute
Subset Description	The description of the experimental variable
Subset Variable Type	The type of experimental variable, e.g., age, strain, gender
GEO Profiles	
Experiment Type	The experiment type, e.g., cDNA, genomic, protein, SAGE
Flag Information	Specific experimental variable flags, e.g., age, strain, gender
Flag Type	Flag types, e.g., rank and value subset effects
GDS Text	DataSet description text
GEO Accession	The GEO accession number (GPLxxx, GSMxxx, GSExxx, GDSxxx)
GEO Description/Title Text	Text provided in the description/title of original records
GI	Mapped GenBank identifier
Gene Description	Gene description, symbol, alias
ID_REF	The unique identifier for a reporter as given on the array
Max Value Rank	The maximum value rank [*]
Max value in profile	The maximum value in profile [*]
Median value in GDS	The median value in DataSet [*]
Median value in profile	The median value in profile [*]
Min Value Rank	The minimum value rank [*]
Min value in profile	The minimum value in profile [*]
Number of Samples	The number of Samples in the DataSet [*]
Organism	The organism from which the Samples were derived
Ranked Standard Deviation	The ranked standard deviation
Reporter Identifier	The identifier for a reporter
Sample Source	The source biological material of the Sample

^a Useful qualifier fields for performing restricted GEO DataSets and GEO Profiles queries.

^{*} indicates possible range operation, e.g., 20:50[Number of Samples] will find DataSets containing 20 to 50 Samples.