GEO: the Gene Expression Omnibus

National Center for Biotechnology Information ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services

GEO Database

The Examination of gene expression using high-throughput methodologies has become very popular in recent years. Techniques such as microarray hybridization and serial analysis of gene expression (SAGE) allow the simultaneous quantification of tens of thousands of gene transcripts. The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes high-throughput gene expression data submitted by the scientific community. GEO currently stores approximately a billion individual gene expression measurements, derived from over 100 organisms, addressing a wide range of biological issues. These huge volumes of data may be effectively explored, queried, and visualized using user-friendly Web-based tools. GEO is accessible at

www.ncbi.nlm.nih.gov/geo

Architecture

Submitters supply their gene expression data in four sections:

Platform: describes the list of features on the array (e.g., cDNAs, oligonucleotides, etc.).

Sample: describes the biological material and the experimental conditions under which the sample was handled, and the abundance measurement of each feature derived from it.

Series: defines a set of related Samples that are considered to be part of an experiment.

Supplementary data: original microarray scan images or raw quantification data.

Sample data are assembled into biologically meaningful and comparable GEO DataSets. DataSet records provide a coherent synopsis about an experiment and form the basis of GEOs data display and analysis tools.

Submissions

An infrastructure is provided in which submitters can supply MIAME-compliant data. There are four ways in which data may be deposited with GEO:

Web deposit: simple, step-by-step, interactive Web forms.

Spreadsheets: Excel spreadsheet templates for easy batch deposit.

SOFT: a plain text, line-based format designed for rapid batch submission.

MINiML: an XML format designed for rapid batch submission.

Complete information about deposit options is provided at:

www.ncbi.nlm.nih.gov/projects/geo/info/submission.html

Data Mining

The data in GEO can be queried using two NCBI Entrez databases:

Entrez GEO-DataSets provides an **experiment-centric** view of the data in GEO. Experiments of interest may be located by searching for attributes such as free text keywords, technology type, author, organism, and experimental variable information. Once a relevant DataSet is identified, that experiment can be further explored for gene expression profiles of interest (Figure 1) using the supplementary tools provided on the DataSet record (Figure 2). Entrez GEO-DataSets is accessible at:

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds

Tools available on the GDS record

—*Cluster heat maps:* A selection of hierarchical and K-means clustering algorithms are provided. Clusters of interest can be selected, enlarged, downloaded, plotted as line charts, or linked directly to Entrez GEO-Profiles.

—*Query subset A vs. B:* This tool assists in the identification of genes that display marked differences in expression level between two specified sets of Samples within a DataSet, as calculated using t-tests or fold difference. Genes that meet the user-defined criteria are presented in Entrez GEO-Profiles.

—*Subset effects:* This feature retrieves all profiles that are flagged as having significant effects with respect to a specific experimental variable, for example 'age' or 'strain'.

Entrez GEO-Profiles provides a **gene-centric** view of the data in GEO. Gene expression profiles (Figure 1) of interest may be located by searching for attributes such as gene name, GenBank accession number, SAGE tag, GEO accession number, description, or profiles flagged as having significant effects with regards to specific experimental variables. Entrez GEO-Profiles is accessible at

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo

Tools available within Entrez GEO-Profiles results page

- *Profile neighbors:* returns a list of genes that show a similar expression pattern within a given DataSet.
- *Sequence neighbors:* retrieves profiles related by nucleotide sequence similarity by BLAST
- *Homolog neighbors:* retrieves profiles of genes belonging to the same HomoloGene group
- *Links:* Links to other NCBI Entrez databases including GenBank, PubMed, Gene, UniGene, OMIM, Homologene, Taxonomy, SAGEMap, and MapViewer.

The data in GEO can also be queried outside of the Entrez databases with

GEO BLAST: The GEO BLAST interface allows users to search for GEO-Profiles of interest based on nucleotide sequence similarity. Additionally, all standard BLAST results display 'E' icons that link directly to GEO-Profiles expression data.

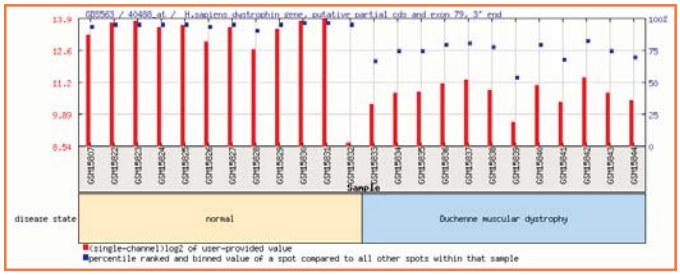

Figure 1: Expression profile of the dystrophin gene in GEO DataSet 563 which examines skeletal muscle biopsies from 12 Duchenne muscular dystrophy patients and 12 unaffected control subjects. Red bars represent gene expression values, blue squares represent intra-sample percentile rank information, providing an indication of the relative expression level of that gene compared to all other genes on the array. Experimental design is reflected in subgroup labels along the bottom of the chart. As expected, the dystophin gene is seen to be expressed at lower levels in Duchenne patients compared with unaffected control subjects.

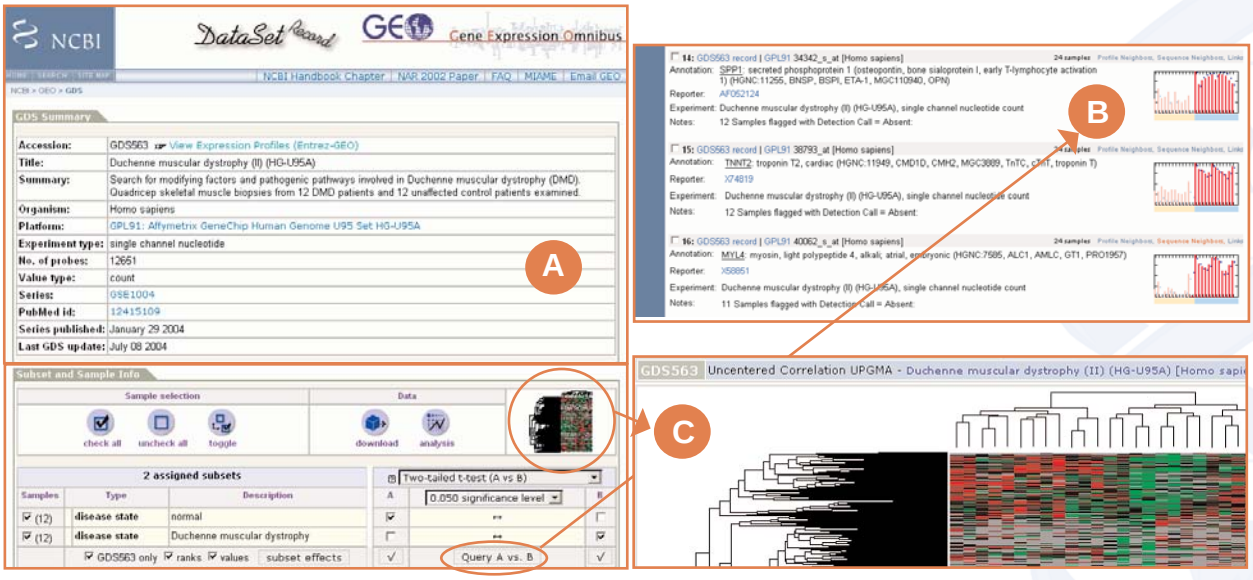

Figure 2: GEO DataSet records contain experiment summary information (A) and access to data mining features such as a 'Query subset A vs. B' statistical tool that identifies profiles of interest (B) based on the experimental conditions and cluster heat maps (C).

Questions relating to GEO submission and GEO query should be sent to: geo@ncbi.nlm.nih.gov

FTP download: All Platform, Sample and Series records, raw data, and GDS value matrices with annotation are available for bulk download via FTP at ftp.ncbi.nih.gov/pub/geo

National Center for Biotechnology Information ■ 8600 Rockville Pike, Building 38A, Room 3S-308
Bethesda, MD 20894 Phone: 301-496-2475 ■ Fax: 301-480-9241 ■ e-mail: info@ncbi.nlm.nih.gov
Web site: www.ncbi.nlm.nih.gov

07/2007