GEO

Gene Expression Omnibus

David Montaner
www.dmontaner.es/materiales
dmontaner@cipf.es

11 March 2013

Microarray Databases I

- GeneNetwork system: Open access standard arrays, exons arrays, and RNA-seq data for genetic analysis (eQTL studies) with analysis.
- UNC modENCODE Microarray database: Nimblegen customer 2.1 million array 6.
- UPSC-BASE: data generated by microarray analysis within Umeå Plant Science Centre (UPSC).
- UPenn RAD database: MIAME compliant public and private studies, associated with ArrayExpress.
- UNC Microarray database: provides the service for microarray data storage, retrieval, analysis, and visualization.

Microarray Databases II

- MUSC database: The database is a repository for DNA microarray data generated by MUSC investigators as well as researchers in the global research community.
- caArray at NCI: Cancer data, prepared for analysis on caBIG.
- ArrayTrack: ArrayTrack hosts both public and private data, including MAQC benchmark data, with integrated analysis tools.
- NCI mAdb: Hosts NCI data with integrated analysis and statistics tools.
- ImmGen database: Open access across all immune system cells; expression data, differential expression, coregulated clusters, regulation.

Microarray Databases III

- Genevestigator database: Gene expression search engine based on manually curated microarray data.
- **Gene Expression Omnibus (GEO)**: NCBI any curated MIAME compliant molecular abundance study.
- **ArrayExpress**: at EBI Any curated MIAME or MINSEQE compliant transcriptomics data.
- Stanford Microarray database: private and published microarray and molecule abundance database.

Source: http://en.wikipedia.org/wiki/Microarray_databases

GEO

Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

MIAME: Minimum Information About a Microarray Experiment)

Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Brazma et. al (2001) Nat Genet. 2001 Dec;29(4):365-71. PMID: 11726920 [PubMed - indexed for MEDLINE]

MIAME

- The raw data for each hybridization (e.g., CEL or GPR files)
- The final normalized data for the set of hybridizations in the study (e.g. gene expression data matrix)
- The essential sample annotation (e.g., compound and dose in a dose response experiment, class)
- The experimental design including sample data relationships (e.g. technical and biological replicates)
- Sufficient annotation of the array (e.g. gene identifiers, genomic coordinates, probe oligonucleotide sequences)
- The essential laboratory and data processing protocols (e.g., what normalization method)

Data Organization in GEO I

Original data (submitted by researchers)

- **Platform** record: summary description of the array template.
  GEO accession number: **GPL**xxx
- **Sample** record: individual sample data (genomic, phenotypic, experimental . . .)
  GEO accession number: **GSM**xxx
- **Series** record: a group of related samples, usually from one experiment or study.
  GEO accession number **GSE**xxx.

Data Organization in GEO II

Curated data (organized by GEO)

- **DataSet** records: a curated collection of *biologically and statistically comparable* samples reassembled by GEO staff form one or several series
  GEO accession number **GDS**xxx.
  For them GEO has data display and analysis tools.
- **Gene Profiles**: measurements for an individual gene across all Samples in a DataSet.

GEO Web Query

Query

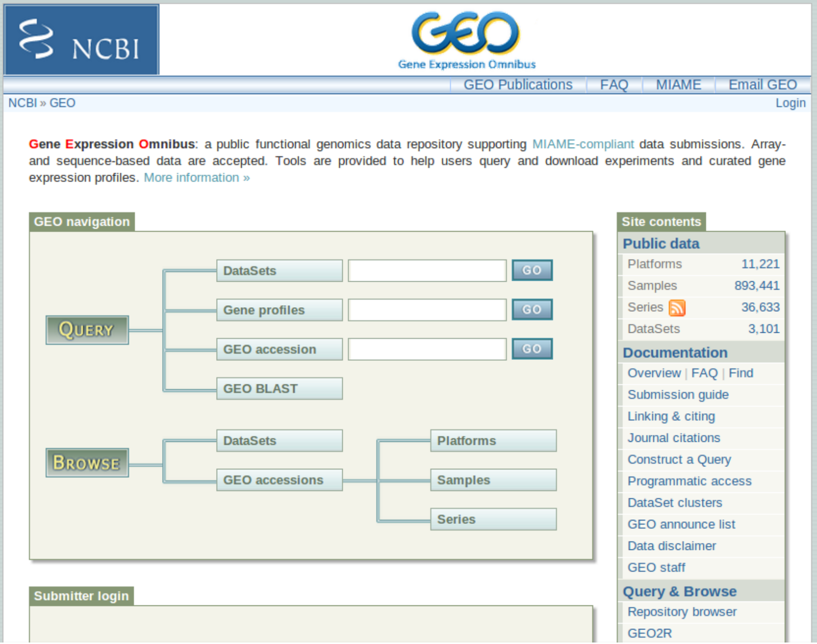
- DataSets: Stores curated gene expression DataSets.
  Search example: *melanoma*
- Gene profiles: Stores individual gene expression profiles from curated DataSets.
  Search example: *melanoma*
- **GEO accession**: Searches GEO Accessions.
  Search example: *GSE37761*

Browser: *nicer interface; exports searches*

- DataSets
  Search example: *melanoma*
- GEO accession: Platforms; Samples; Series
  Search example: *melanoma*

`http://www.ncbi.nlm.nih.gov/geo/`

GEO Web



Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information »

Site contents

Public data

Platforms	11,221
Samples	893,441
Series	36,633
DataSets	3,101

Documentation

Overview | FAQ | Find

Submission guide

Linking & citing

Journal citations

Construct a Query

Programmatic access

DataSet clusters

GEO announce list

Data disclaimer

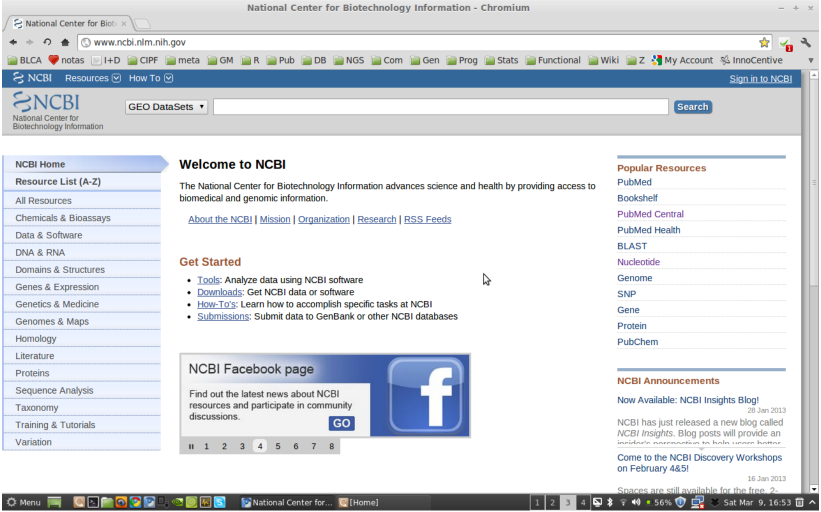
GEO staff

Query & Browse

Repository browser

GEO2R

GEO at NCBI



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

NCBI Facebook page

Find out the latest news about NCBI resources and participate in community discussions.

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

Now Available: NCBI Insights Blog!

28 Jan 2013

NCBI has just released a new blog called *NCBI Insights*. Blog posts will provide an insider's perspective to help users better

Come to the NCBI Discovery Workshops on February 4&5!

16 Jan 2013

Spaces are still available for the free, 2-

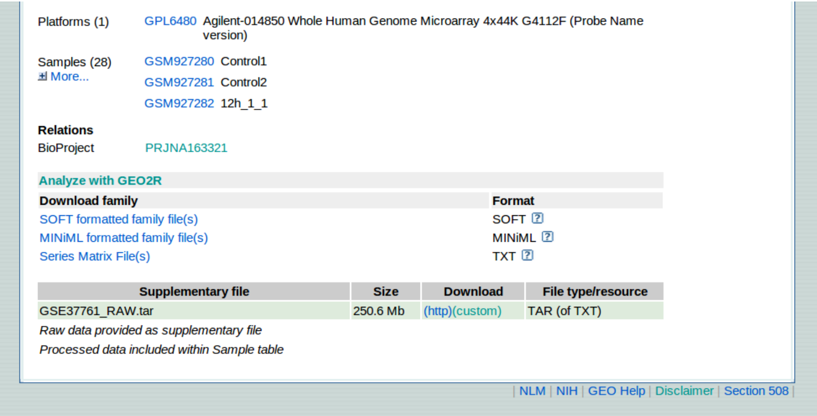
Usual searches

- Search for a GEO accession . . . form a publication.
- Search for example data for a particular platform
- Keyword
- Date . . .

Download GEO data

- Links on Series records are provided at the foot of each GEO Series record web page. Ex. GSE37761 web.
- FTP download: `ftp://ftp.ncbi.nlm.nih.gov/geo/`. Ex. `ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE37nnn/GSE37761/`
- Programmatic access to GEO: server-side programs to retrieve data; can be used with a fixed URL syntax.

Series Data Formats I



Platforms (1)	GPL6480	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version)
Samples (28) ⊞ More...	GSM927280	Control1
	GSM927281	Control2
	GSM927282	12h_1_1

Relations

BioProject	PRJNA163321

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE37761_RAW.tar	250.6 Mb	(http)(custom)	TAR (of TXT)

Raw data provided as supplementary file

Processed data included within Sample table

Series Data Formats II

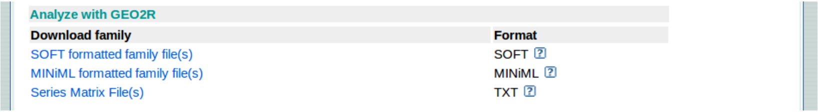
- SOFT formatted family file(s): complete data and metadata (gene information) in a single file.
- MINiML formatted family file(s): complete data and metadata in separated files.
- Series Matrix File(s): complete data in a tab delimited matrix; no metadata information.

- Supplementary files: usually raw data.

We generally use the *Series Matrix* format and may be the *platform* file within the *MINiML* folder.

GEO internal tools

GEO2R: simple analysis for GEO Series or DataSets



Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

- explore *Value distribution*: box-plot and summary statistics.
- explore single gene expression *Profile graph*
- perform a differential expression analysis to compare two or more groups of Samples.
- clustering (just for DataSets)

Bioconductor Packages

http://www.bioconductor.org/packages/release/

- GEOmetadb: A compilation of metadata from NCBI GEO.
- GEOsubmission: Prepares microarray data for GEO submission.
- GEOquery: **Get data** from NCBI Gene Expression Omnibus.

References

- http://www.ncbi.nlm.nih.gov/geo/info/
- http://en.wikipedia.org/wiki/Microarray_databases
- http://en.wikipedia.org/wiki/MIAME