



---

*Accessing Public Data  
(from NCBI) Using Bioconductor*

---

Sean Davis, M.D., Ph.D.  
Genetics Branch, Center for Cancer Research  
National Cancer Institute  
National Institutes of Health

# National Institutes of Health



# Objectives

- Navigate NCBI GEO website
- Understand NCBI GEO data *entities* and relationships between them
- Use GEOquery package to import data from NCBI GEO into R
- Convert GEOquery data structures into R data structures
- Use GEOmetadb to find data in NCBI GEO
- Know relationship between NCBI GEO and NCBI SRA
- Understand how to use the SRADB package to query SRA metadata
- Use SRADB package to control the Integrated Genome Viewer (IGV) from R



You must  
explain your  
problem clearly





# MIAME-compliant Data

- The raw data for each hybridization (e.g., CEL or GPR files)
- The final processed (normalized) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
- The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
- The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates)
- Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

# What's in GEO

- Gene expression profiling by microarray or next-generation sequencing
- Non-coding RNA profiling by microarray or next-generation sequencing
- Chromatin immunoprecipitation (ChIP) profiling by microarray or next-generation sequencing
- Genome methylation profiling by microarray or next-generation sequencing
- Genome variation profiling by array (arrayCGH)
- SNP arrays
- Serial Analysis of Gene Expression (SAGE)
- Protein arrays

# GEO data entities

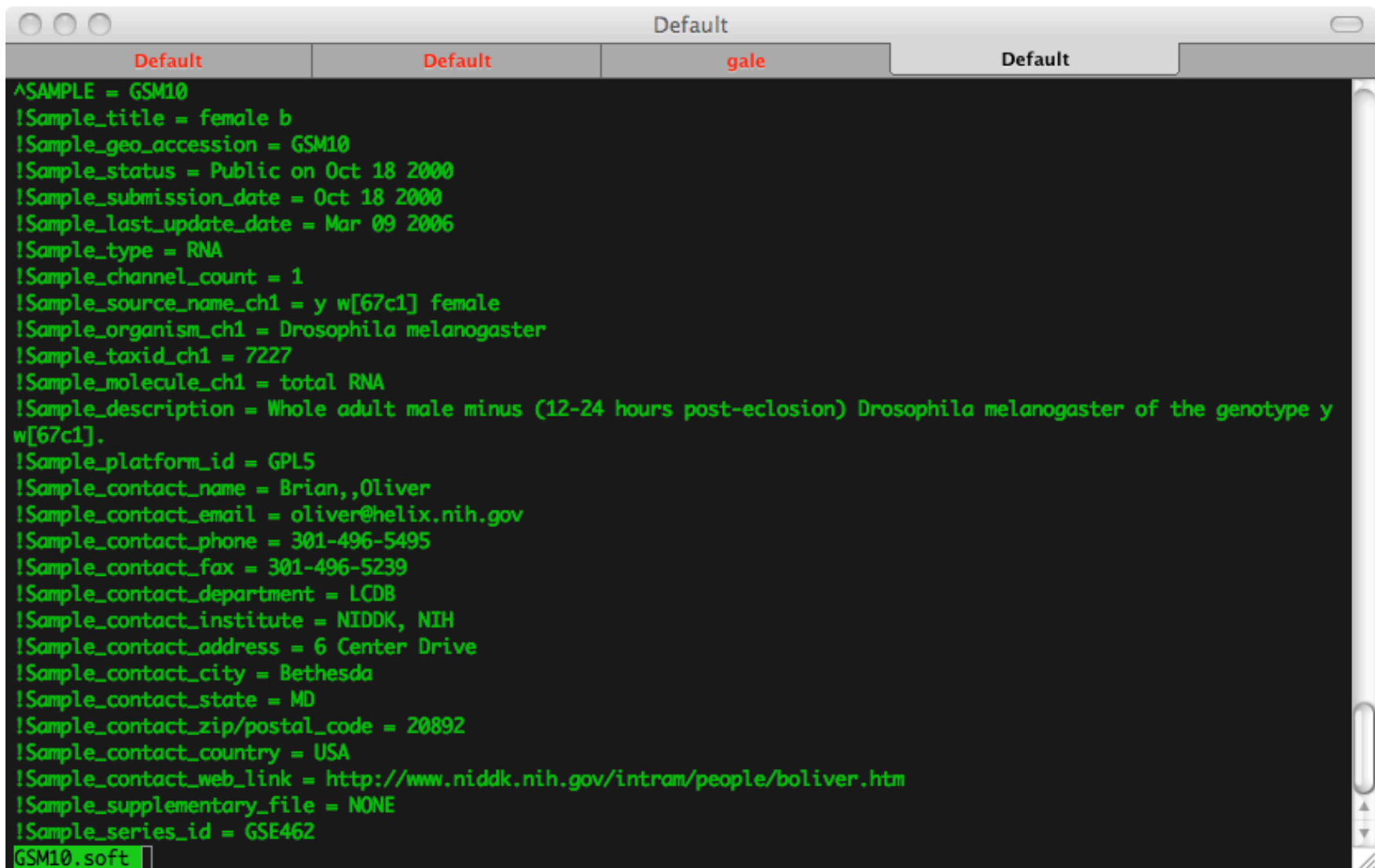
- GEO Samples (GSM)
- GEO Platform (GPL)
- GEO Series (GSE)
  - Collections of related GSM and GPL along with *free-text* study metadata
  - Two data “flavors”
    - GSE
    - GSEMatrix
- GEO Dataset (GDS)
  - Only data entity that is curated by NCBI GEO staff
  - Typically, samples are divided into statistically and biologically relevant groups upon which we can compute



# NCBI GEO website

- <http://www.ncbi.nlm.nih.gov/geo/>

# GEO SOFT Format



The image shows a screenshot of a text editor window with a dark background and green text. The window title is "Default". The text content is as follows:

```
^SAMPLE = GSM10
!Sample_title = female b
!Sample_geo_accession = GSM10
!Sample_status = Public on Oct 18 2000
!Sample_submission_date = Oct 18 2000
!Sample_last_update_date = Mar 09 2006
!Sample_type = RNA
!Sample_channel_count = 1
!Sample_source_name_ch1 = y w[67c1] female
!Sample_organism_ch1 = Drosophila melanogaster
!Sample_taxid_ch1 = 7227
!Sample_molecule_ch1 = total RNA
!Sample_description = Whole adult male minus (12-24 hours post-eclosion) Drosophila melanogaster of the genotype y w[67c1].
!Sample_platform_id = GPL5
!Sample_contact_name = Brian,,Oliver
!Sample_contact_email = oliver@helix.nih.gov
!Sample_contact_phone = 301-496-5495
!Sample_contact_fax = 301-496-5239
!Sample_contact_department = LCDB
!Sample_contact_institute = NIDDK, NIH
!Sample_contact_address = 6 Center Drive
!Sample_contact_city = Bethesda
!Sample_contact_state = MD
!Sample_contact_zip/postal_code = 20892
!Sample_contact_country = USA
!Sample_contact_web_link = http://www.nidk.nih.gov/intram/people/boliver.htm
!Sample_supplementary_file = NONE
!Sample_series_id = GSE462
GSM10.soft
```

# GEO SOFT Format

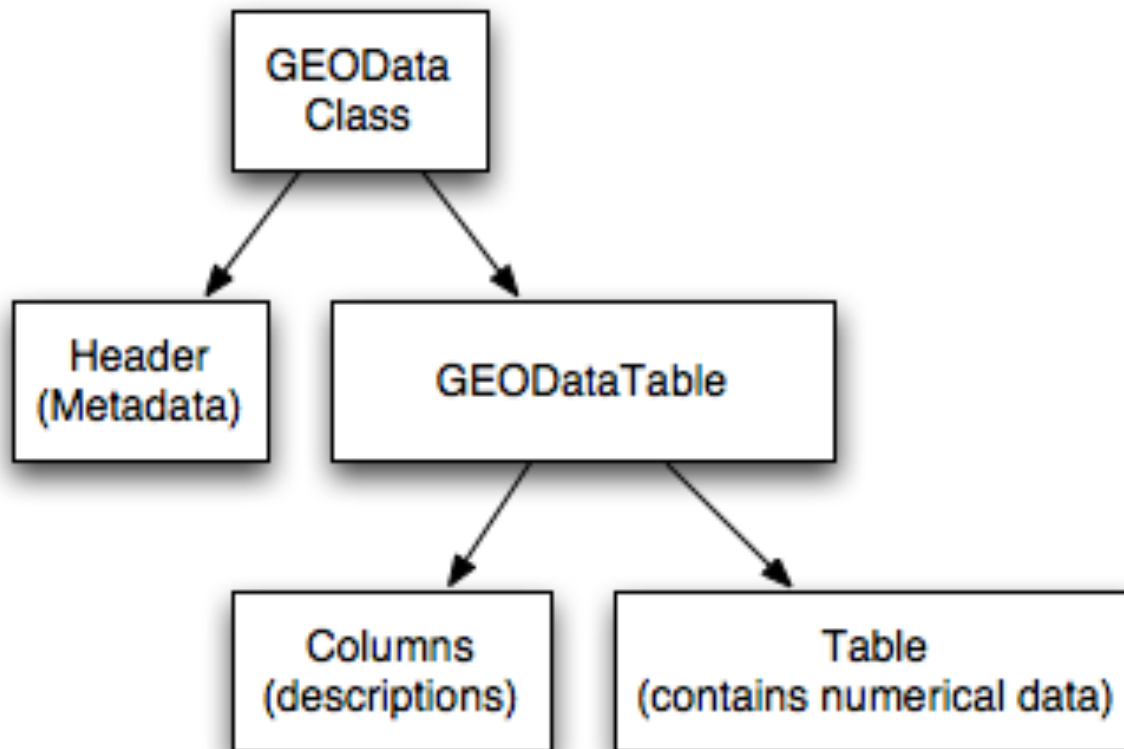
```
Default
Default      gale      Default
!Sample_data_row_count = 3456
#ID_REF =
#SIGNAL_RAW = raw signal
#BKD_FORM =
#NORM_FORM =
#BKD_RAW = raw background
#NORM_VALUE = normalization value
#CONST = constant value
#VALUE =
!sample_table_begin
ID_REF  SIGNAL_RAW  BKD_FORM  NORM_FORM  BKD_RAW  NORM_VALUE  CONST  VALUE
1       4486.49 0      0      3379.57875  23337.5396  39542  55845.4489
2       3482.51 0      0      3379.57875  23337.5396  39542  41058.05142
3       3812.39 0      0      3379.57875  23337.5396  39542  45916.78036
4       3257.56 1      0      3379.57875  23337.5396  39542  37744.81305
5       5436.91 0      0      3379.57875  23337.5396  39542  69843.97308
6       4053.44 0      0      3379.57875  23337.5396  39542  49467.15204
7       3729.72 0      0      3379.57875  23337.5396  39542  44699.15237
8       7344.6  0      0      3379.57875  23337.5396  39542  97941.91357
9       4128.13 0      0      3379.57875  23337.5396  39542  50567.24439
10      8273.2  0      0      3379.57875  23337.5396  39542  111619.0558
11      3513.57 0      0      3379.57875  23337.5396  39542  41515.52723
12      4171.29 0      0      3379.57875  23337.5396  39542  51202.9384
13      3848.21 0      0      3379.57875  23337.5396  39542  46444.36515
14      3242.18 0      0      3379.57875  23337.5396  39542  37518.28446
15      3674.06 0      0      3379.57875  23337.5396  39542  43879.34865
16      3802.43 0      0      3379.57875  23337.5396  39542  45770.08174
17      3234.09 0      0      3379.57875  23337.5396  39542  37399.12865
18      4584.82 0      0      3379.57875  23337.5396  39542  57293.72953
:
```



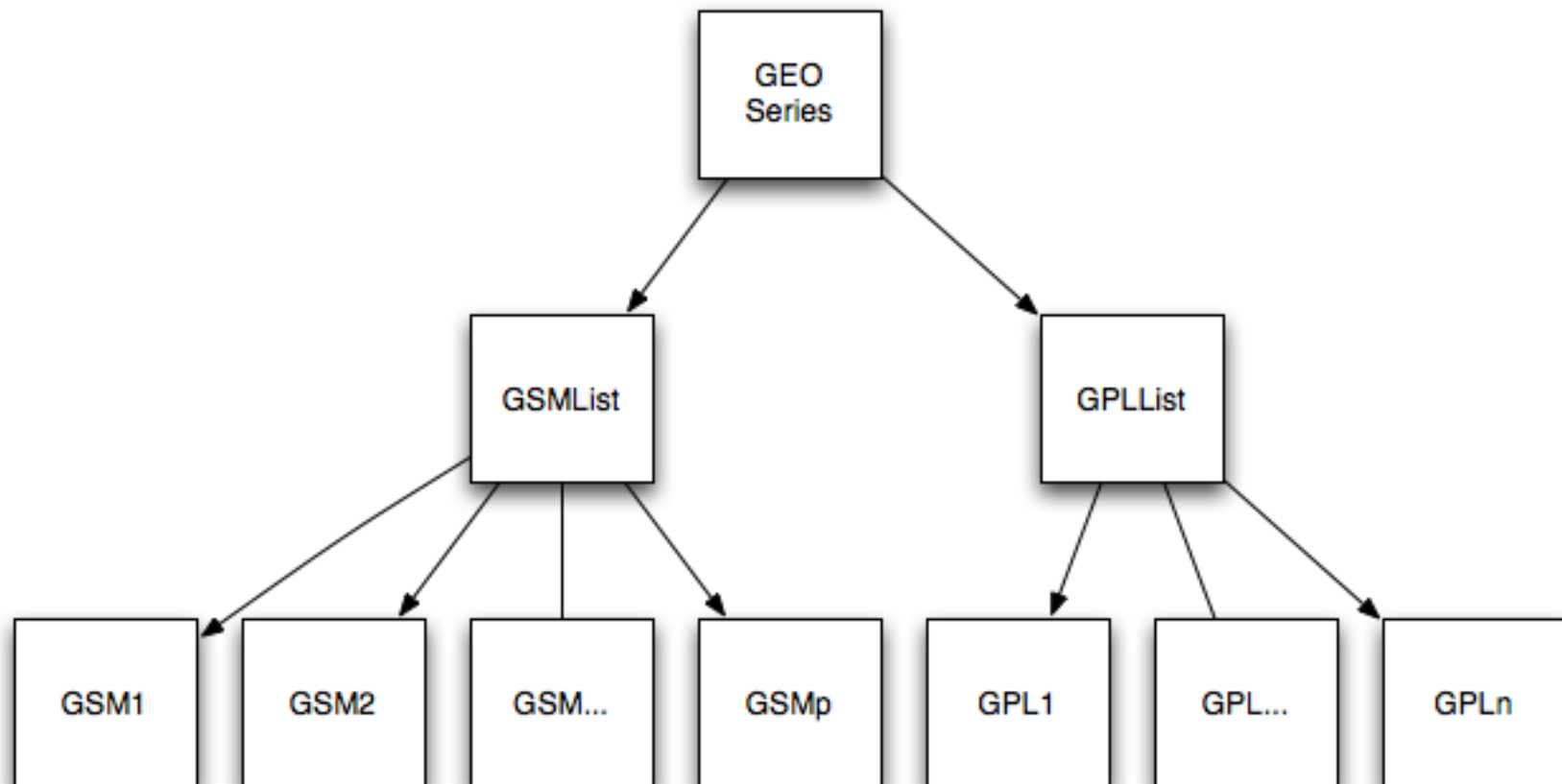
# GEOquery

- Singular goal: *Get data from NCBI GEO and parse into R objects*
- Secondary goals:
  - Get supplemental files from NCBI GEO
  - Allow large-scale data mining by doing all-of-the-above in a lossless manner

# GEOquery Data Structures (GSM, GPL, GDS)



# GEOquery Data Structures (GSE)





# GEOquery Walkthrough

- <http://watson.nci.nih.gov/~sdavis/>
  - Click on Tutorials and then on “Accessing public data....”
- Download the R script for cut-and-paste to follow along

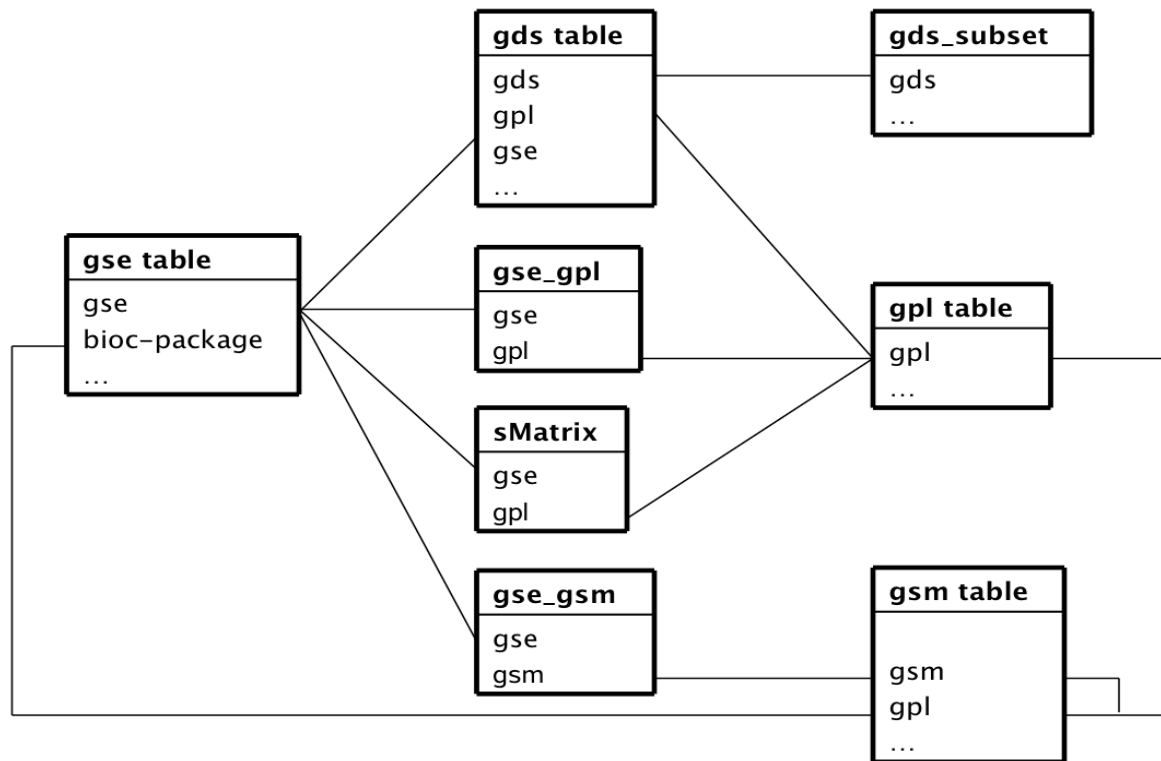
# GEOmetadb

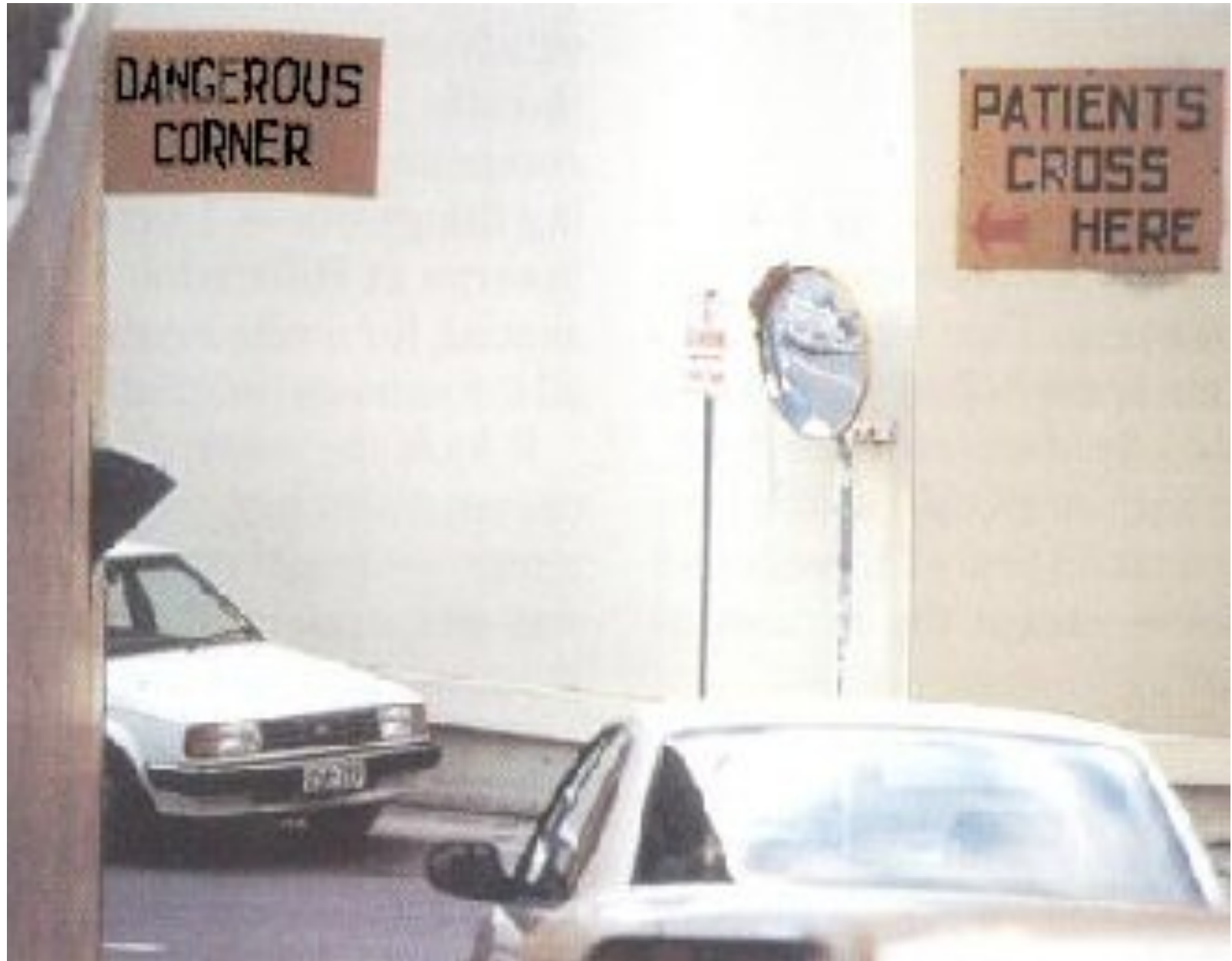
- Finding data in NCBI GEO can be challenging
- Some investigators need large-scale, computable access to GEO metadata
- Given one GEO entity type, it may be useful to find all relationships with other entity types (eg., find all hgu133a arrays in GEO)

# GEOmetadb

- What is GEOmetadb?
  - A Bioconductor package that offers an alternative to eUtils data mining of GEO metadata
  - A SQLite database which stores all the GEO metadata in a relational format for easy querying, particularly in bulk
- What is GEOmetadb NOT?
  - We have not attempted to alter or standardize in any way the data from GEO

# GEOmetadb Schema





# GEOmetadb Walkthrough

# SRADB

- Similar goals to GEOmetadb, but with SRA data
- Data for SRA (ENA, DRA) are mirrored, but exact policies are a bit unclear (to me)
- Accessing to the data also provided by SRADB package, but further processing (SRA SDK) now needed to get FASTQ files

# SRAdb and NCBI SRA

Recently, NCBI announced that due to budget constraints, it would be discontinuing its Sequence Read Archive (SRA) and Trace Archive repositories for high-throughput sequence data. However, NIH has since committed interim funding for SRA in its current form until October 1, 2011. In addition, NCBI has been working with staff from other NIH Institutes and NIH grantees to develop an approach to continue archiving a widely used subset of next generation sequencing data after October 1, 2011.

We now plan to continue handling sequencing data associated with:

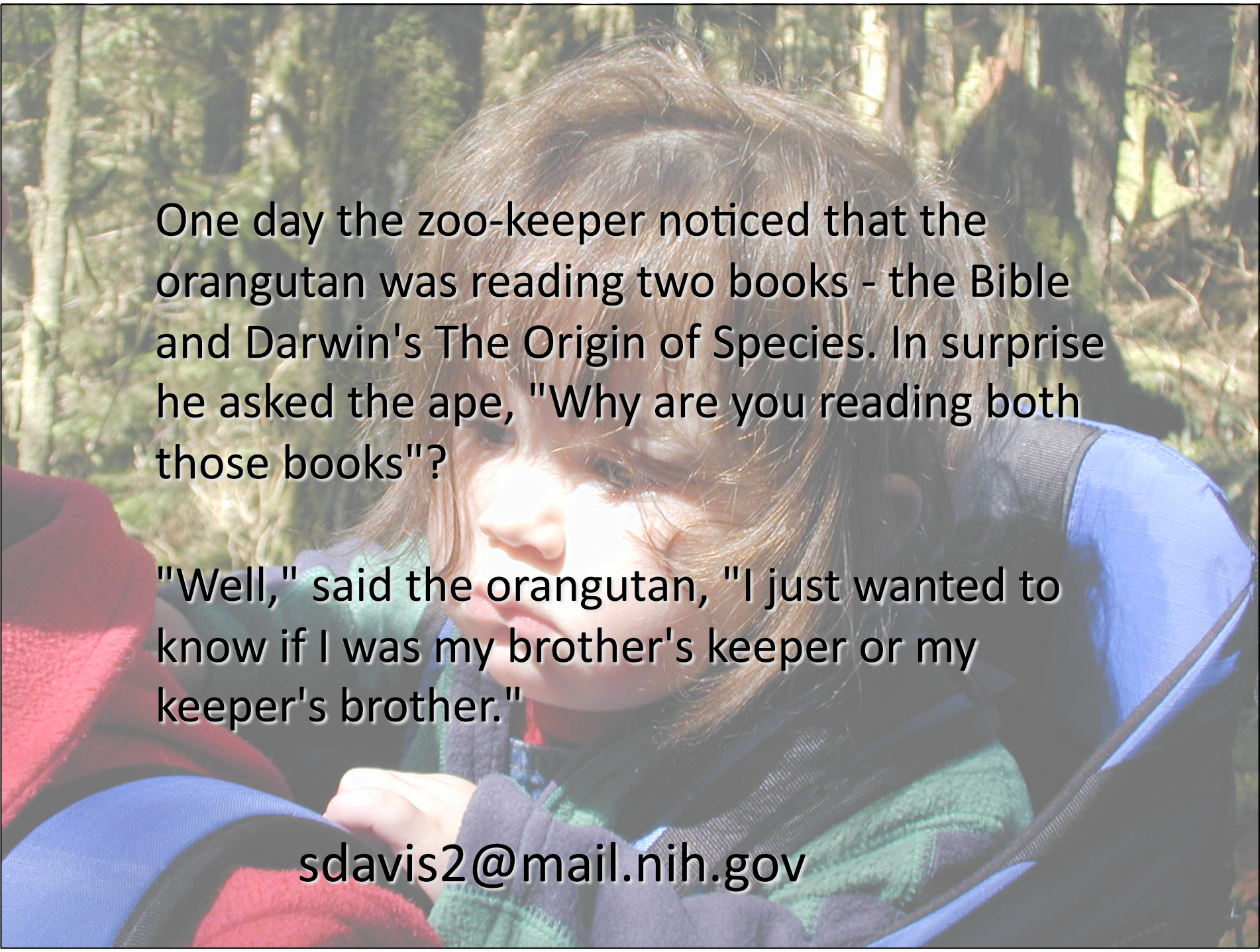
- RNA-Seq, ChIP-Seq, and epigenomic data that are submitted to GEO
- Genomic and Transcriptomic assemblies that are submitted to GenBank
- 16S ribosomal RNA data associated with metagenomics that are submitted to GenBank

In addition, NCBI will continue to provide access to existing SRA and Trace Archive data for the foreseeable future. NCBI is also continuing to discuss with NIH Institutes approaches for handling other next-generation sequencing data associated with specific large-scale studies.



# SRADB and IGV Walkthrough

- Download and start IGV:
  - <http://www.broadinstitute.org/software/igv/download>



One day the zoo-keeper noticed that the orangutan was reading two books - the Bible and Darwin's The Origin of Species. In surprise he asked the ape, "Why are you reading both those books"?

"Well," said the orangutan, "I just wanted to know if I was my brother's keeper or my keeper's brother."

[sdavis2@mail.nih.gov](mailto:sdavis2@mail.nih.gov)