Scope

Biological studies in the post-genome era often produce large volumes of complex and high-throughput datasets of multiple types. For example, a single study may address topics of genomic sequencing, transcriptome, epigenome and genotype/phenotype association, or it could be further divided into multiple sub-projects, each focused on a narrower field. The different characteristics of datasets generated by a complex study require that they be deposited in different databases at NCBI, such as GenBank, SRA (Sequence Read Archive), dbGaP (database of Genotype and Phenotype), Gene Expression Omnibus (GEO), or others. Registering a study in the BioProject database \[1, 2\] generates a central record that provides a clear overview of the scope of that study and functions as a primary key to link its divergent datasets and allow easy data access. Registration in BioProject is mandatory for new genome-scale studies with primary data submitted to the International Nucleotide Sequence Database Collaboration (INSDC). The BioProject database entries are closely related to the corresponding entries in the Genome database and can be used as a conduit to the genomic sequence data.

Accessing and registering projects

The BioProject homepage (www.ncbi.nlm.nih.gov/bioproject/) provides access to records for registered projects through a text search (A) or through a browser using Project attributes (B). Records from other NCBI databases also link to Bio-Project records. Help documentation (C) provides guidance for registration of projects and help in using the resource to find records. The Submission (D) link accesses NCBI’s submission portal for the registration of new projects in BioProject. Registering a project requires credentials from an MyNCBI, NIH or eRA login.

Searching with “Chinchilla [organism]” (E) retrieves a list of study projects for this organism, which can be further filtered using facets in the left column (F).
Using the Advanced page

The Advanced page provides access to indexed fields and terms indexed within them. It also provides a search builder function to assist the construction of complex query terms with proper field limits to help retrieve records fit specified criteria.

In this page, clicking the index field displays available fields in a pull-down menu (A), which can be selected (highlighted) for use as a field limit. Clicking the “Show index list” link (B) adds terms indexed under the selected field. A selected term automatically appears in the search box (C) above. The “Add to history” link (D) searches the database with terms in the search box and adds the result to the history list. Unlocking the search box using the “Edit” link (E) allows custom input. This example combines two existing searches with AND. Clicking a number in the “Items found” (F) retrieves the results. The example above highlights three very important and informative fields: organism, filter, and properties.

Displaying a project record

Contents displayed in project records differ for umbrella projects and data-containing primary projects. An umbrella project provides an overview serves as a centralized entry point where all sub-projects under it can be readily retrieved through links within the body of the record. The example umbrella project (shown to the right) is a human ENCODE project. The top section provides a summary on the scope of the project and the project type (G). The number of available links to datasets in other databases (H) reflects the scale of the project. Links in the “Navigate” panel (I) allows the navigation to the parent and sister projects.
Actual experimental data are accessible through primary project links listed in the bottom panel (A).

A primary project has a simpler display and provides direct links to datasets generated by the project. In the example, the data links are shown in the “Project Data” table (B) as well as the more traditional Entrez links in the “Related information” section (C) to the right. The record also provides links to the genome record and related bioproject entries in the “Navigate” panel (D).
Retrieving linked data from other databases

Links within a primary project make the retrieval of experimental data readily accessible. For example, the sequence reads from different experimental next generation sequencing runs are available through the SRA link (A). The samples used in these experiments can be found through the BioSample link (B). Finally, the genome record can be displayed through the Genome link (C) above the "Navigate" panel. More information on the genome display is available in an article in the NCBI Newsletter [3].

References
3. NCBI Newsletter on the release of the new Genome database: http://1.usa.gov/y88y85