

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

MeshLinker

Permalink

<https://escholarship.org/uc/item/3md246zi>

Author

Lee, Roy,

Publication Date

2004

Peer reviewed|Thesis/dissertation

MeshLinker: an automated web-based tool for organizing genes under MeSH

by

Roy Lee, M.D.

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco



Date

University Librarian

**Copyright © 2004
By
Roy Lee, M.D.**

Dedication

To my outstanding research advisors, Dr's Tom Ferrin and Patsy Babbitt, from the UCSF BMI program. While most PI's were wary of letting a master's student pick a bioinformatics project, the two of you extended a golden opportunity for me to do so, and not only did I learn a lot, but I am glad I made the decision I did. I cannot thank both of you enough for taking me under your wings. I hope to make both of you proud with my future accomplishments.

To Conrad Huang and the rest of the CGL gang – you guys are truly the Jedi masters of Python and SQL. I'd like to think I'm a much better developer now, instead of the newbie I was when I first started.

To my friend Corey Adams. I deeply value your honest friendship, and wish the best to you for the rest of your graduate school career. I promise to get back into shape soon!

To my mother and sister, who are the world to me. And finally, to my beagle, Bishop, of 14.5 years, who suddenly passed away during the final few days of this manuscript's preparation – you will be sorely missed. Hats off to you for the wonderful years you were with me, from college, through medical school, my active duty years in the Air Force, and these last couple years at UCSF.

Cheers to Noriko, Rie, and Junko.

Abstract

There has been a recent explosion on the amount of genetic sequence data in recent years, and much of this data lacks any correlation to clinical significance, despite a clear interest in making gene-disease associations in the academic research communities. MeshLinker, a web application for organizing large sequence datasets under a clinical ontology, was built to address that particular interest. MeshLinker establishes gene-disease associations in an automated fashion. It also provides web links to outside databases such as UniGene and OMIM, and categorizes each gene under the MeSH hierarchy by parsing PubMed abstracts for MeSH headings. Lastly, it provides a searchable/browsable web interface to view this information.

Table of Contents

Dedication	iii
Abstract	iv
Table of Contents	v
List of Tables.....	vi
1. Introduction	1
BayGenomics.....	2
Motivation.....	4
2. Specific Aims	6
3. Methods.....	7
Data Retrieval	7
Choice of Ontology.....	9
4. System Architecture and Software.....	15
5. Draft of a Manuscript to be Submitted.....	16
6. Validation and Testing	22
7. Discussion	25
8. Conclusion.....	28
9. References	30
10. Appendices	33
Appendix A: Overview of the NCBI databases and tools	33
GenBank	33
LocusLink.....	34
UniGene.....	34
HomoloGene.....	35
OMIM.....	35
PubMed.....	36
Entrez Utilities.....	36
11. Tables	37
12. Figures/Illustrations.....	43

List of Tables

Table 1: List of the original 11 PGA centers:	37
Table 2: PGA Topics Under Investigation:.....	37
Table 3: Comparison of sample clinical ontologies	38
Table 4: List of Python scripts and modules from MeshLinker.....	39
Table 5: MeSH Categorizations of Baygenomics Genes	40
Table 6: BayGenomics cell lines with known phenotypes	41
Table 7: List of German Gene Trap Consortium cell lines with known phenotypes.....	42

List of Figures/Illustrations

Figure 1: Programs for Genomic Applications	43
Figure 2: BayGenomics.....	44
Figure 3: Screenshot of a BayGenomics data access page.....	45
Figure 4: Screenshot of a BayGenomics annotation page.....	46
Figure 5: Flowchart for data retrieval	47
Figure 6: Database schema.....	48
Figure 7: MeshLinker title page	49
Figure 8: Example browser page displaying MeSH subcategory for “Diseases”	50
Figure 9: Browser page with gene annotations	51
Figure 10: Example search results page using query of “parkinson’s disease”	52

1. Introduction

In the realm of informatics, there has often been a culture clash between basic and clinical scientists. There are those that strongly feel that the disciplines of bioinformatics and medical informatics should be separate academic programs. On the other hand, one can make the argument that their overall goals are similar: impact our understanding of biology and physiology that can be harnessed to impact patients in a positive way.

Altman recently made the argument that the two different fields of bioinformatics and medical informatics can have powerful synergies and technology transfers between them [1]. One of the ultimate goals of both bioinformatics and clinical informatics is to have robust computational models of physiology that will enable us to model, store, retrieve, and analyze the effects, on patients, of disease, medications, and the environment. Each of these two disciplines approach this same goal from opposite ends of the basic science-clinical science spectrum. Technologies involving bioinformatics often emphasize basic biological data such as sequences, structures, pathways, and genetic networks. Technologies involving medical informatics include knowledge representation, data mining, automated diagnosis, and information retrieval.

There are a wide variety of bioinformatics databases in use today, each with their own approach to the kind of data presented. GenBank [2] is a large genetic sequence database, containing both nucleotide and protein sequences for all species, housed at the National Center for Biotechnology Information (NCBI). LocusLink¹ [3, 4] provides a single query interface to curated sequence and descriptive information about genetic loci. UniGene [3, 5] is a database that clusters all sequences into non-redundant, gene-oriented

clusters. OMIM [6] (Online Mendelian Inheritance in Man) is a database of human genes and genetic disorders, originally developed at Johns Hopkins, and now maintained at NCBI. It contains textual information and references - there are links to MEDLINE, sequence records in Entrez [3], and other resources at NCBI and elsewhere. More information on these databases can be found in Appendix A of this manuscript.

BayGenomics

On September 30, 2000, the National Heart, Lung, and Blood Institute (NHLBI) launched the Programs for Genomic Applications [7] (PGA) (Figure 1), providing a \$40 million grant divided amongst 11 centers across the U.S. (Table 1). This program is a major initiative to advance functional genomic research related to heart, lung, blood, and sleep health and disorders.

The UCSF PGA is divided into 9 different components, all covering different aspects of the project, such as gene-trapping, bioinformatics gene identification and database management, in-situ hybridization, and microarrays, and education (Table 2). Specifically, the UCSF PGA involves the BayGenomics web resource [8] (Figure 2), which is a database housing thousands of mouse sequences isolated by experimental gene-trapping [9].

Gene-trapping [10] is an experimental technique which essentially allows isolation of a random area of DNA and blocking its expression in the embryonic stem cell. This “knockout” is propagated through generations of cell division and multiplying,

¹ LocusLink is due to be replaced by the newer Entrez Gene, and will cease operation on March 1, 2005. Because LocusLink was used for this thesis' prototype application, this manuscript will still refer to it.

and if it does not cause an embryonically lethal mutation, it can result in a living mouse model that lacks the expression of the gene in question.

Component 1 utilizes gene-trap vectors to inactivate approximately 2500 genes per year in mouse embryonic stem (ES) cells. To date, approximately 10,000 cell lines have been trapped. All trapped ES cells are posted on the BayGenomics website and distributed for a nominal charge by the NIH/NCRR-sponsored Mutant Mouse Regional Resource Centers (MMRRC) [11] to the scientific community for the purpose of producing knockout mice.

Component 2 is intimately involved with bioinformatics techniques and uses automated computational approaches to identify gene-trap sequences to known, full-length sequences in the non-redundant (NR) GenBank database. We have linked 92% of our 10,000 BayGenomics cell lines to approximately 17,000 genes, which include multiple genes corresponding to a specific cell line (Figure 3). By identifying these sequences and providing useful annotation to each gene-trap sequence (such as providing mappings to outside databases like LocusLink, GenBank, and MGI/JAX [12]), BayGenomics provides a very useful service to the scientific community, which can acquire any of the various cell lines for conducting research on knockout mice (Figure 4).

One of the ultimate goals of BayGenomics is to assess which of the ES cell clones are involved in cardiopulmonary development and common cardiopulmonary diseases, thereby increasing the value of the resource to investigators.

Motivation

The usability of mice as excellent animal models can not be overstated. Over the past century, the mouse has developed into the premier mammalian model system for genetic research. Scientists from a wide range of biomedical fields have gravitated to the mouse because of its close genetic and physiological similarities to humans. Although yeasts, worms, and flies are excellent models for studying the cell cycle and many developmental processes, mice are far better tools for probing the immune, endocrine, nervous, cardiovascular, skeletal and other complex physiological systems that mammals share. Like humans and many other mammals, mice naturally develop diseases that affect these systems, including cancer, atherosclerosis, hypertension, diabetes, osteoporosis and glaucoma. Adding to the mouse's appeal as a model for biomedical research is the animal's relatively low cost of maintenance and its ability to quickly multiply, reproducing as often as every nine weeks [13].

On the other hand, BayGenomics was originally focused on the needs of research scientists, and hence is rather gene and nucleotide sequence oriented. Though it may be useful to the scientific gene knockout community, the usability to clinicians not as well-versed in bioinformatics may be limited. Work involved with this thesis project was performed to provide a prototype strategy for linking gene identities for the genetraps to available clinical information principally via MeSH and OMIM.

We feel that development of a tool that can extend the knowledge domain of BayGenomics closer to the clinical realm can help facilitate new types of queries not currently possible with BayGenomics – and possibly open up new avenues of investigation for researchers. As an example, it currently is not possible to search for

UCSF LIBRARY
AVENUE 350M

specific genes with a disease-oriented query such as “show me all of your genes that are linked to cardiovascular disease”. Rather, the user is limited to searching by categories such as GenBank accession number, gene symbol name, and MGI/JAX number.

First, a connection from our PGA mouse genes to OMIM human entries would provide an extremely useful addition to BayGenomics. The utilization of OMIM information helps bridge the gap between our mouse bioinformatics site and clinical research, by providing relevant connections between our gene-trap sequences and human genetic disease. It also provides one method to determine which of our genes are related to cardiopulmonary and/or sleep disorders - one of the overall PGA goals.

Furthermore, once these valuable connections to human diseases are made, this information can be presented to the user by utilizing a clinical vocabulary in a search engine. By doing so, we can facilitate sophisticated clinically minded queries that are not currently possible on BayGenomics. The ability to search the BayGenomics database by disease name in addition to accession number or cell line number would allow clinical researchers to quickly hone in on relevant genes and cell lines, and thereby facilitate the creation of appropriate mouse models for their research activities.

UCSF LIBRARY
MAY 17 1990

2. Specific Aims

- (1) *Develop a set of Python modules to translate and integrate information between disparate genomic databases, and to map genes into a clinically oriented ontology.*
- (2) *Develop a database schema to represent relevant relationships and store data obtained from parsing modules.*
- (3) *Develop a web-based front end, utilizing browser and search engine functionality that can be integrated into the BayGenomics website.*

UCSF LIBRARY

3. Methods

(1) Develop a set of Python modules to translate and integrate information between disparate genomic databases, and to map genes into a clinically oriented ontology.

A wide variety of biological databases exist, such as GenBank, SWISS-PROT, UniGene, and LocusLink. Their features and annotation are often rich, but difficult to integrate with each other due to different number/identification schemes. Often times, data referring to the same actual gene will be referred by different ID numbers in different databases – for example, the gene for “adenomatosis polyposis coli” has a GenBank accession number of NM_007462, a GenBank GI number of 6680692, a LocusLink ID of 11789, and a MGI/JAX ID of 88039. It is fairly trivial to “jump across” different databases manually with a web-browser, given a single gene – however it is not when dealing with a large set of 10,000 different gene accession numbers. An automated system is clearly desired.

The Python programming language [14] was used to write a series of automated scripts to obtain equivalent identifier numbers from some of these different databases, given the starting point of a single GenBank accession number. In addition, given a single gene, all associated PubMed are retrieved. Doing so allows retrieval of entire PubMed abstracts, which contain Medical Subject Headings (MeSH) [15].

Data Retrieval

Starting with the GenBank accession number, the modules perform two different paths for information retrieval from the internet (Figure 5): one path retrieves UniGene,

AMERICAN
UNIVERSITY
LIBRARY

HomoloGene [3], and OMIM data, and the other retrieves data from PubMed, to include MeSH headings. The first path obtains the accession number, retrieves the LocusLink number, makes the UniGene connection, and then determines the most probable human homolog to this mouse gene, via HomoloGene. Once the human homolog is determined, then the scripts will attempt to retrieve a matching OMIM entry, if one exists for it. The second path retrieves the GenBank information for the particular gene accession number, and retrieves all PubMed abstracts associated with that GenBank page. Using the PubMed abstracts, the scripts parse out all MeSH headings from them.

Whenever possible, XML information was retrieved from NCBI, using their e-Utilities (also known as e-Utills) interface [16], which provides their data in encapsulated XML tags. XML is often preferable to HTML and/or plain text, because many websites often change the format of their displayed output – when this occurs, it will often break the parser that was written. XML encapsulated fields distinctly identify various data fields, and are changed much less often. This approach is superior to basing a parser on HTML documents. Essentially, XML parsers can specifically look for encapsulated tags such as <Gene_id>, while with HTML there is no direct way to identify which data on the page is really the gene ID. HTML pages are often parsed using regular expression text matching, and this approach often fails when websites changes how their pages are displayed.

Initially, NCLEVER4.0 [17] was used to retrieve MEDLINE abstracts. NCLEVER4 is an old command line program, dating back to the early 1990's, that connects over the network to NCBI's Entrez database, and can run batch jobs involving many queries. The source code is available on the internet and easily compiled on a

Linux workstation. NCLEVER was used in a recent paper that also performed MeSH term matching, similar to what was done with this project [18]. However, with the advent of e-Utils, the step involving NCLEVER became unnecessary, as e-Utils was used to retrieve the same information, and much more, using the same interface.

The entire retrieval process took approximately 4-5 days of continuous activity using 17,000 gene accession numbers. The process could potentially be done much faster, but delays were used between each connection to NCBI to comply with their request for not flooding their network server – on their user requirements page, they state that users should not make on average more than one request per 3 seconds.

A modified version of Conrad Huang's e-Utilities library was utilized for XML retrieval from e-Utils (personal communication with Conrad Huang, November 2003). These scripts were written in Python. In the case of e-Utils not capable of retrieving needed data, such as UniGene and HomoloGene names, then a direct HTML query was used to query the NCBI Entrez site, and retrieve XML data embedded within HTML. XML parsing was accomplished with Sam Schreiber's Multiloader2 package (personal communication with Sam Schreiber, June 2004), also written in Python.

Choice of Ontology

There were a number of clinical ontologies to pick from: SNOMED, Medical Subject Headings (MeSH), International Classification of Diseases (ICD-9), and Unified Medical Language System (UMLS) [19] are a few examples. These four ontologies are compared in Table 3.

The smallest of the four, MeSH is used to organize and index scientific literature in Medline and PubMed. The MeSH ontology is fairly small, at 25 megabytes of text

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

information and contains over 22,000 concepts. It contains approximately 300,000 synonyms for those concepts as well. An average of 10-15 MeSH indexing terms (also known as descriptors) are assigned to each scientific abstract by professional indexers at NCBI – these indexers manually assign keywords after reading the article. Incorporating MeSH into a process system as described above is a natural extension of the data gathering process – MeSH headings are easily obtained for all scientific articles while searching with the PubMed system.

SNOMED [20] and ICD-9 [21] are much larger ontologies, each having at least 10 times the number of concepts as MeSH. SNOMED was developed by pathologists and is primarily used in a clinical setting. ICD-9 is familiar to many physicians as it is used for billing and insurance purposes. However, there is no simple, easy way to map genes to these ontologies, as they are clinically based and are not a part of the downloaded PubMed entries.

UMLS is intended to be the all-inclusive ontology, incorporating over 100 source vocabularies such as MeSH and SNOMED, and containing over 1 million concepts. It provides a metathesaurus that links similar concepts between these different ontologies. UMLS is very large with over 20 gigabytes of uncompressed information, and comes on a DVD disc due to its size. Its size makes it more complete than the other ontologies. It is also very complex because it semantically links all the different concepts from different ontologies together. Overall, its size and complexity makes it very unwieldy and extremely challenging to use. Users of UMLS will often prune the UMLS ontology into their own customized (and smaller) version. For example, for the purpose of

categorizing mouse knockout genes, it would probably be safe to exclude all concept trees related to botany and/or medicolegal affairs.

Of worthy mention, the Gene Ontology (GO) [22] is a widely used and very well known ontology. However, it is aimed at the basic science spectrum more than clinical. It would not be possible to search for disease phenotypes such as myocardial infarction or Parkinson's Disease. GO is centered around three structured, controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions.

Because of the simplicity and accessibility of using MeSH, we chose to use it for this project over other ontologies such as UMLS.

The various Python module files written for this project are summarized in Table 4.

(2) Develop a database schema to represent relevant relationships and store data obtained from parsing modules.

Information and data management is one of the major themes of both medical informatics and bioinformatics. The use of tens of thousands of genes can easily result in many megabytes worth of annotation to be stored and retrieved. Though flat files are easier to use and implement at first, a relational database management system (RDBMS) is clearly the better choice with this amount of information, due to scalability and performance issues.

MySQL 3.23 [23] was chosen as the RDBMS platform due to a variety of reasons: open source, fast performance, no need for a dedicated database administrator, and portable among most operating system platforms.

MySQL is characterized as a free, fast, reliable open source relational database [24]. It lacks some sophistication and facilities, but it has an active development team and, as it goes from release to release, more capabilities are added. At certain times there will be a trade-off between speed and capabilities, and the MySQL team intend to keep their database engine fast and reliable.

By comparison, PostgreSQL [25] is another open source relational database management system. It conforms to the SQL standards much better than MySQL and runs on a wide variety of platforms, but the extra feature set slows and complicates its use. Oracle is the leading commercial RDMS; it is highly flexible, runs on many platforms and has a full and sophisticated feature set. Because of its highly tunable nature, an Oracle database administrator (DBA) needs to be well and heavily trained. In comparison, MySQL does not require a dedicated DBA. Microsoft SQL Server runs only on Windows platforms, which excludes its use in applications where there is a need to run on Linux or Unix.

Figure 6 depicts the database schema that was developed for this project. Overall, the schema can be divided into three major sections. The first section comprises of tables already in use by the production BayGenomics database server. These tables contain the gene accession numbers and LocusLink numbers. The two other sections contain information retrieved and processed by the MeshLinker scripts: one containing the UniGene/HomoloGene/OMIM information, and the other containing the PubMed

UNIVERSITY OF TORONTO LIBRARY

abstract information (journal, title, authors, etc.) as well as MeSH headings derived from the abstracts. Lookup tables are also included which link MeSH terms to MeSH numbers, and also MeSH synonym terms. This information is obtained and parsed directly from downloaded files made available on the MeSH homepage at:

<http://www.nlm.nih.gov/mesh/meshhome.html>

(3) Develop a web-based front end, utilizing browser and search engine functionality that can be integrated into the BayGenomics website.

The preceding step generates a massive amount of information: for the 17,000 BayGenomics genes used as input, over 23,000 entries were deposited in the abstracts table, and over 56,000 entries deposited in the mesh table. Therefore, an efficient interface is necessary to query and view this large collection of information. The natural solution was to build a web-based front end system, using the Apache webserver application [26] and written in the Python programming language. In order to resemble many common search engines in existence today, this required building two things: (1) a browser hierarchy, and (2) search engine functionality.

The browser (Figure 7) involves dynamically generated web pages and allows users to traverse up and down through the tree structure of a MeSH (Figure 8), viewing the relationships between different MeSH terms, determining the number of BayGenomics genes categorized under each term, and viewing each gene's various annotations and links to other databases (Figure 9).

Representing the MeSH tree in a browser hierarchy required using the MeSH numbers associated with each term (e.g. "C04.682" = "Neoplasms, Radiation-Induced").

Starting at the top level MeSH number prefix (e.g., “C”), the hyperlinks lead to subsequent branches until a leaf node is reached and one cannot drill down any deeper in that particular branch (“C”, “C04”, “C04.682”, “C04.682.512”, etc.)

Matching genes to individual MeSH terms was made possible by using the associated PubMed abstracts in MEDLARS format and parsing out the “MH” (MeSH Headings) lines from it. Given a particular gene and the PubMed abstracts corresponding to it, each abstract contained on average 10 to 15 MeSH headings. Each of these obtained headings were then deposited into the MeshLinker database, with information leading back to the gene in question. In this fashion, we were able to link genes to their associated MeSH heading categories.

Search engine functionality was incorporated to allow direct queries such as “which genes fall under the term cardiomyopathy?” Each downloaded MeSH record for a term also contains various synonyms for that term, labeled as “ENTRY”. For example, the entry for the MeSH descriptor “Parkinson Disease” contains a number of synonym terms: “Idiopathic Parkinson’s Disease”, “Lewy Body Parkinson’s Disease”, “Parkinson Disease, Idiopathic”, “Parkinson’s Disease”, “Parkinson’s Disease, Lewy Body”. Any query using these terms will result in the same MeSH category shown on the web browser screen (Figure 10).

It should be noted that many common terms for particular diseases and conditions are not recognized as MeSH terms or synonyms. A query of “MI” or “heart attack” results in a “search not found” error page – they do not result in the page for “Myocardial Infarction”, which is the official MeSH term for that particular concept.

4. System Architecture and Software

MeshLinker was initially developed on an Intel Pentium4 workstation under the Debian Linux distribution, running at 3 GHz CPU with 1 GB of RAM memory. Open source software used in conjunction with MeshLinker are:

- MySQL 3.23 relational database management system
- Apache 1.3 webserver
- Python 2.3 programming language

MeshLinker was later ported to the Socrates cluster server housed at the UCSF Resource for Biocomputing, Visualization and Informatics (RBVI) [27], for increased performance. Because the operating system was also UNIX-based (Hewlett Packard Tru64), porting to the new system was fairly trivial. MySQL, Apache, and Python were already available on the new system. Socrates is based on Hewlett Packard's (HP) AlphaServer family of computers and includes a 32-processor GS1280 and four 4-processor ES45s.

5. Draft of a Manuscript to be Submitted

To be submitted to Bioinformatics journal under the Applications Note section.

MeshLinker: an automated web-based tool for organizing genes under MeSH.

Roy Lee, Patricia Babbitt, and Thomas Ferrin.

ABSTRACT

Summary: We developed MeshLinker, an automated web application for organizing large sequence datasets under a clinical ontology. MeshLinker establishes links to outside databases such as UniGene and OMIM, categorizes each gene under the MeSH hierarchy by parsing PubMed abstracts for MeSH headings, and provides a searchable/browsable web interface to view this information.

Availability: MeshLinker is available at <http://baygenomics.ucsf.edu/mesh>

Contact: tef@cgl.ucsf.edu

INTRODUCTION

The recent explosion in genomic research has resulted in huge and unwieldy genetic sequence datasets with often useful molecular annotation, but still lacking disease and developmental context at the larger organism level. MeshLinker is an automated system for categorizing large and/or custom sets of genes within a searchable/browsable medical subject heading (MeSH) hierarchy [15], and providing useful links to relevant outside

databases. Researchers can utilize this software in conjunction with their own collection of genes to generate UniGene [5], OMIM [6], PubMed [2], and MeSH linkages, and display this information in an easy to use web browser interface. We present as an example the use of MeshLinker as a way to help organize the collection of mouse gene knockouts generated by the BayGenomics consortium and to quickly view their disease associations in the biomedical literature.

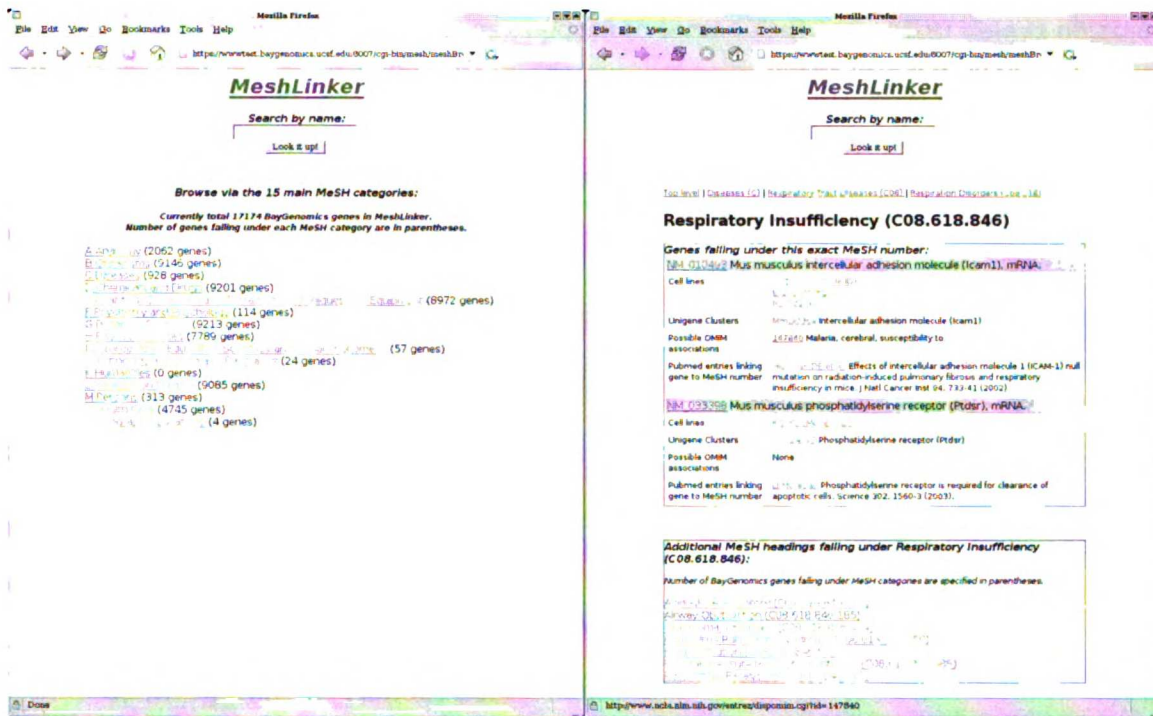


Figure 1. Sample screenshots. From left: (1) Front page of website, displaying the top level MeSH tree. One may either browse the MeSH tree or simply type in a search query. (2) A detailed display of the Respiratory Insufficiency category, along with the 2 genes matching that exact MeSH number. URL links to GenBank, BayGenomics, UniGene, OMIM, and PubMed abstracts are given for each gene.

SYSTEM AND METHODS

The MeshLinker system does primarily four things: (1) categorizes genes within the MeSH hierarchy using the assigned MeSH headings from their associated PubMed abstracts, (2) retrieves additional information and related annotation from other outside databases, (3) organizes all the data in a relational database, and (4) provides a web browser front end for querying this database. All data is retrieved over the internet from NCBI using e-Utils [16] or Entrez, as opposed to local CD-ROM. Most data is obtained as XML formatted text, and processed through a custom parser called Multiloader2 (personal communication with Sam Schreiber, June 2004).

Starting with a gene accession number, the gene's GenBank entry is parsed for all PubMed ID numbers, which in turn allow the retrieval and parsing of PubMed abstracts for MeSH headings and other annotation. Other annotation that is retrieved, stored, and displayed are UniGene clusters and OMIM entries. If the gene in question belongs to a non-human organism, an attempt is made to link to OMIM by searching for a possible human homolog from the HomoloGene database [3].

A series of Apache CGI scripts provide the web interface to this information. The interface is minimalist and the style similar to that seen with the search/browser functions seen at various search engine websites. A simple text search box is provided, along with a clickable tree to allow traversing through the MeSH hierarchy. The search box will consider alternative spellings/synonyms for a given query by utilizing the set of terms

AMNH
FROM

within a particular MeSH entry. If a particular gene has been linked to a MeSH category, then a wide variety of information and outside links (UniGene, OMIM, GenBank, etc.) are displayed in the browser application. Furthermore, any associated PubMed URL links to abstracts are provided for the particular gene.

MeshLinker is written in the Python programming language and developed on a Debian Linux server. It interfaces with a MySQL database and an Apache web server.

RESULTS AND DISCUSSION

This system represents a powerful application that can add clinical relevance to a custom set of gene data. Its main advantage is utilizing any set of GenBank gene accession numbers, and viewing their MeSH associations, through dynamic webpages. We have applied this to link data from mouse gene knockouts to disease terms associated with MeSH. The mouse gene knockouts are generated and maintained through the BayGenomics consortium, funded by the NHLBI Program in Genomic Applications [9]. Though we have used it in the context of categorizing our mouse gene knockouts, MeshLinker could also be used to categorize sets of genes for other species, as well as microarray probe sets.

Overall, there are 17,174 unique genes linked to BayGenomics cell line sequences used in the current instance of MeshLinker. Of those, 9220 genes are categorized somewhere in the MeSH hierarchy – there are many GenBank entries for these 17,174 genes without

19

links to articles in PubMed. Out of these 9220 genes, 928 genes are categorized under the "C" MeSH subcategory for "Diseases". Overall, there are 324 unique genes linked to entries in OMIM.

For these 928 genes with some kind of linkage to a MeSH category, 45% of them are associated with nervous system diseases, 38% with immunologic disease, and 33% with neoplasms. It is important to note that a gene may be linked to multiple MeSH categories.

Future efforts are directed at improving these numbers, which we expect as more knockout genes are obtained and identified through the BayGenomics project. Future methods for mapping our cell lines and genes to MeSH will likely improve these numbers as well.

MeSH contains approximately 22,500 concepts, and though we incorporate MeSH synonyms, the search engine still was not able to find any results for the query "heart attack" – this is simply because that query is not a MeSH term, while "myocardial infarction" is. The use of higher level, more comprehensive ontologies such as the Unified Medical Language System (UMLS) [19] and SNOMED [20] will likely improve the quality of searching. Because MeSH has been integrated within UMLS, we are investigating the possibility of incorporating UMLS technology to resolve non-MeSH term queries to proper MeSH numbers by first determining the correct UMLS concept ID

UNIVERSITY OF MICHIGAN LIBRARY

6. Validation and Testing

For these 928 genes with some kind of linkage to a MeSH category, 45% of them are associated with nervous system diseases, 38% with immunologic disease, and 33% with neoplasms (Table 5). It is important to note that a gene may be linked to multiple MeSH categories – for example, gene accession NM_008543 is *Mus musculus MAD homolog 7 (Drosophila) (Madh7), mRNA*, and in the literature has been linked to different MeSH categories such as diabetic nephropathies, squamous cell carcinoma, dermatitis, scleroderma, and hepatocellular carcinoma.

Using MeSH descriptors for certain genes as a means of establishing some sort of biomedical categorization can be misleading. Often, linked MeSH descriptors are extremely vague and not too informative. To take the above example gene, NM_008543 is also categorized under Animals (B01), Proteins (D12.776), Tooth (A14.549.167.860), and Air (G03.230.300.100.150). None of these categories give the user any useful information.

On the other hand, there are genes that appear to be categorized accurately and informatively. There are a few genes with OMIM linkages that have the MeSH descriptor terms mentioned somewhere in their OMIM record. As an example, NM_009680 is *Mus musculus adaptor-related protein complex AP-3, beta 1 subunit (Ap3b1), mRNA*. Through the process of establishing homology connections (via Homologene) and ultimately an OMIM entry, this mouse gene was linked to a human OMIM record with a title of *Hermansky-Pudlak syndrome*, which is a form of oculocutaneous albinism. The MeSH descriptor that this gene was categorized under was

Albinism, Oculocutaneous (C11.270.040.545). It is therefore reasonable to conclude that NM_009680 was placed by the automated system into a correct MeSH category.

It is possible to develop further automation to detect genes which have the same terms between the OMIM entry and the MeSH descriptor. Such automation would give us the ability to quickly determine which genes were categorized with a high degree of accuracy. However, it would not be trivial to design and implement – for example, performing a simple text match of *oculocutaneous albinism* would fail against *Hermansky-Pudlak syndrome*. One would also consider having the automated program view all the words in the OMIM entry, and not only the OMIM title. In the above example, the term *oculocutaneous albinism* appeared in the body text of the OMIM entry, and not the title itself.

Two gene sets were used in an attempt to validate the process of MeSH correlation – these were gene sets with known phenotypes and/or diseases. The first was a list of 5 BayGenomics cell lines with known disease phenotype, supplied by Dean Sheppard, MD of the Gladstone Institute (Table 6). The second was a list of several genetrapp sequences from the German Gene Trap Consortium (GGTC) [28], also with known phenotypes (Table 7).

The five sample BayGenomics cell lines were all identified to one or more GenBank accession numbers – however, none of these genes were linked to any MeSH descriptors related to the supplied phenotypes from Dr. Sheppard. Looking at each case individually, it was discovered that the automated MeshLinker process was not necessarily at fault. For all of the five unlinked genes, the supplied phenotype names did not appear as the appropriate MeSH descriptor in the PubMed article abstract. As an

example, cell line RRK003, described by Dr. Sheppard as involved with acute lung injury, is identified to GenBank accession numbers NM_010442, M33203, and X56826. Examining the PubMed abstracts and their associated MeSH headings, none of the three genes had any assigned MeSH headings related to the term *acute lung injury* or its synonyms (e.g. acute respiratory distress syndrome).

Another reason why performing a search for these genes failed was because in one case, the supplied phenotype was not a MeSH term. Cell line RRS565 is linked to hypoplastic lungs, according to the list. However, the term *hypoplastic lungs* is not an official MeSH descriptor or synonym, and therefore did not result in any matches when that query was entered.

The list of sequences from GGTC did not easily correspond to our BayGenomics cell lines and/or genes, due to different cell line labeling and use of non-standardized gene symbol names. Because of this, it was not a simple matter to use that information to validate the MeshLinker system.

7. Discussion

MeshLinker appears to work quite well for finding genes and cell lines, given a disease of interest. However, as seen with the two example data sets, it does not appear to work well the other way around; we tend to have poor results finding MeSH correlations, given known genes. Though the mentioned reasons for not matching known phenotype to MeSH descriptor are understandable, we would certainly like better results.

There are a number of possible directions to take in the future with the MeshLinker system. The first is to improve the quality and accuracy of these automated gene-disease connections. The second is to continue widening the potential user base for this application.

Many opportunities to improve upon the quality and accuracy of the gene-disease connections exist, but none are trivial, and all require completely different methods for establishing these connections. We currently utilize MeSH as the ontology of choice. Though it is very easy to use and implement for a prototype such as the one we have created here with MeshLinker, there are some serious deficiencies with relying solely on MeSH, as seen with the problems with validation.

One possibility is to utilize UMLS/Snomed in addition to MeSH. UMLS is much larger and more of a “complete” ontology to use for these purposes. MeSH is already incorporated into UMLS, and the use of semantic relationships in UMLS would be very advantageous. One big problem with using MeSH is the limited synonyms available to use – while a system based on MeSH would have difficulty understanding what the term

“heart attack” is, UMLS likely would not. The largest drawback of using UMLS is its very large size and unwieldiness, as it comes as a package involving several CD-ROM’s. However there are new tools for UMLS these days, namely *MetamorphoSys* [29], which may be used to significantly “prune down” the size of the ontology that is worked with. As an example, it would be possible (and recommended) to remove entries dealing with botany and create a sub-set of the overall UMLS to work with.

There are also other completely different methods for establishing gene-disease connections. Natural language processing (NLP) methods have been used in the past to automatically extract gene names and gene and protein interactions from text. A recent paper [30] described a statistical algorithm that can swiftly identify from the literature, sets of genes known to be associated with given diseases – it offers a comprehensive way to treat alias symbols, a statistical method for computing the relevance of the gene to the query, and a novel way to disambiguate gene symbols from other abbreviations. However, due to the complexity of different gene symbol conventions, resulting in many non-standardized symbols, this method was beyond the scope of this initial project. For example, the acronym ER could be confused with either *Emergency Room* or *Estrogen Receptor*, the latter of which uses the alias symbol ER for the gene ESR1.

Genetic association studies also represent potential avenues of gene-disease associations. Most common diseases are complex disease traits, with multiple genetic and environmental components contributing to susceptibility. It has been proposed that common genetic variants, including single nucleotide polymorphisms (SNP’s), influence susceptibility to common disease [31]. By determining the genotype of these variants in individuals with disease and in unaffected controls, these polymorphisms could be tested

for association with susceptibility to a variety of diseases. These “association studies” usually have a case-control design, where frequencies of the alleles or genotypes at the site of interest are compared in populations of cases and controls; a higher frequency in cases is taken as evidence that the allele or genotype is associated with increased risk of disease. NIH is currently working on a new database to consolidate the data from all known genetic association studies – this is the Genetic Association Database (GAD) [32], located at <http://geneticassociationdb.nih.gov>. It aims to collect, standardize, and archive genetic association study data and to make it easily accessible to the scientific community.

Lastly, widening the usage of MeshLinker is an obvious direction. Though we currently use mouse knockout sequences as the basis for our database, it not need be limited to them. It would be quite interesting to use microarray probe sets, or large gene sets from other species such as yeast and rat.

8. Conclusion

The recent explosion in biomedical data over the last several years underscores the importance and need for the ability to make gene-disease connections. A hybrid approach, combining both bioinformatics and medical informatics principles, would be helpful in tackling this important issue. By understanding the role a particular gene (or lack thereof) plays in normal development or pathological disease, we can advance our knowledge and understand science better. We can harness the power of this newfound knowledge and help bring new cures to diseases by targeting our research efforts at the relevant genes of interest. Hundreds of new genetic sequences are deposited in our public databases every day, and so the increased need is obvious for such automated data mining tools as MeshLinker. Furthermore, as more literature gets published in the journals, more information in the form of MeSH headings will get deposited into MeshLinker as well.

We have demonstrated MeshLinker to be a useful initial approach to helping researchers find genes they would want to investigate. It is a downloadable package that is easily modified to the end-user's needs, utilizing the user's custom list of gene accession numbers – whether microarray probe sets, or large genetrapp sequence sets such as with BayGenomics.

We found making the gene-disease connection via MeSH easily implemented but prone to many misleading or uninformative gene-disease connections. Nevertheless, it served well for a prototype tool such as MeshLinker. The perfect solution has not been discovered yet, and is considered the “holy grail” by some out in industry. It will likely involve other ontologies such as UMLS and Snomed, and incorporate other methods of

automated gene-disease connections, such as statistical-based literature methods. Though a totally automated, 100% accurate solution is still far away, it is still beneficial to have a system that automates mostly everything (retrieve information from NCBI, deposit into database, and conduct queries), and leaves the end user to simply verify and validate the results of his/her search.

9. References

1. Altman RB: **The interactions between clinical informatics and bioinformatics: a case study.** *J Am Med Inform Assoc* 2000, **7**(5):439-443.
2. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA *et al*: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31**(1):28-33.
3. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E *et al*: **Database resources of the National Center for Biotechnology Information: update.** *Nucleic Acids Res* 2004, **32 Database issue**:D35-40.
4. Maglott DR, Katz KS, Sicotte H, Pruitt KD: **NCBI's LocusLink and RefSeq.** *Nucleic Acids Res* 2000, **28**(1):126-128.
5. Pontius JU, Schuler GD: **UniGene: a unified view of the transcriptome.** In: *The NCBI Handbook*. Bethesda: National Center for Biotechnology Information; 2003.
6. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30**(1):52-55.
7. **Programs for Genomic Applications.** [<http://www.nhlbi.nih.gov/resources/pgaf/>]
8. **BayGenomics.** [<http://www.baygenomics.ucsf.edu/>]
9. Stryke D, Kawamoto M, Huang CC, Johns SJ, King LA, Harper CA, Meng EC, Lee RE, Yee A, L'Italien L *et al*: **BayGenomics: a resource of insertional mutations in mouse embryonic stem cells.** *Nucleic Acids Res* 2003, **31**(1):278-281.
10. Stanford WL, Cohn JB, Cordes SP: **Gene-trap mutagenesis: past, present and beyond.** *Nat Rev Genet* 2001, **2**(10):756-768.
11. **Mutant Mouse Regional Resource Centers.** [<http://www.mmrrc.org/>]

12. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT: **The Mouse Genome Database (MGD): the model organism database for the laboratory mouse.** *Nucleic Acids Res* 2002, **30**(1):113-115.
13. Spencer G. **Background on Mouse as a Model Organism.** National Human Genome Research Institute, 2002 [<http://www.genome.gov/10005834>]
14. **Python Programming Language.** [<http://www.python.org/>]
15. Nelson S, Johnston, D, and Humphreys, BL.: **Relationships in medical subject headings.** In: *Relationships in the organization of knowledge.* Edited by Green Ba. New York: Kluwer Academic Publishers; 2001: 171-184.
16. **NCBI e-utils.** [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html]
17. Rioux PA, Gilbert WA, Littlejohn TG: **A portable search engine and browser for the Entrez database.** *J Comput Biol* 1994, **1**(4):293-295.
18. Nagashima T, Silva DG, Petrovsky N, Socha LA, Suzuki H, Saito R, Kasukawa T, Kurochkin IV, Konagaya A, Schonbach C: **Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS.** *Genome Res* 2003, **13**(6B):1520-1533.
19. Lindberg C: **The Unified Medical Language System (UMLS) of the National Library of Medicine.** *J Am Med Rec Assoc* 1990, **61**(5):40-42.
20. **SNOMED Clinical Terms (SNOMED CT).** [<http://www.snomed.org>]
21. **International Classification of Diseases, Ninth Revision (ICD-9).** [<http://www.cdc.gov/nchs/icd9.htm>]
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology.** The Gene Ontology Consortium. *Nat Genet* 2000, **25**(1):25-29.
23. **MySQL: The World's Most Popular Open Source Database.** [<http://www.mysql.org>]
24. Ellis G. **Advantages and disadvantages of MySQL.** 2004 [<http://www.wellho.net/forum/The-MySQL-Relational-Database/Advantages-and-disadvantages-of-MySQL.html>]
25. **PostgreSQL.** [<http://www.postgresql.org>]

26. **Apache Software Foundation.** [<http://www.apache.org>]
27. **UC San Francisco Resource for Biocomputing, Visualization and Informatics (RBVI).** [<http://www.rbvi.ucsf.edu>]
28. **German Gene Trap Consortium (GGTC).** [<http://tikus.gsf.de>]
29. **UMLS MetamorphoSys Fact Sheet.**
[<http://www.nlm.nih.gov/pubs/factsheets/umlsmetamorph.html>]
30. Adamic L.A. WD, Huberman B.A., Adar E.: **A Literature Based Method for Identifying Gene-Disease Connections.** In: *IEEE Computer Society Bioinformatics Conference: 2002; Stanford University; 2002: 109-117.*
31. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies.** *Genet Med* 2002, **4**(2):45-61.
32. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36**(5):431-432.

10. Appendices

Appendix A: Overview of the NCBI databases and tools

GenBank

GenBank is a database of nucleotide sequences from >130,000 organisms. Records that are annotated with coding region (CDS) features also include amino acid translations. GenBank belongs to an international collaboration of sequence databases (described below), which also includes EMBL and DDBJ. GenBank is updated daily in NCBI search systems, and a full release is issued on the FTP site approximately the 15th of every February, April, June, August, October, and December. It contains all the data present in GenBank as of the cutoff date specified in the release notes. The FTP site also provides daily cumulative and non-cumulative update files.

Each GenBank includes, for example, information about accession number formats, sequence identifiers (GI number and accession.version), a listing of GenBank divisions, and more. Each record describes some commonly annotated biological features, such as CDS, and provides links to documents that list and define the complete set of biological features that can be annotated on sequence records.

GenBank is accessible using a web browser through Entrez Nucleotides, and can be accessed using a variety of command-line utilities, as well as through an e-mail interface. An option to download the GenBank full release and updates via FTP is also available.

For more information about GenBank, go to
<http://www.ncbi.nih.gov/Sitemap/index.html#GenBank>

LocusLink

LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci. LocusLink issues a stable ID for each locus and presents information on official nomenclature, aliases, sequence accession numbers, phenotypes, EC numbers, OMIM numbers, UniGene clusters, map information, and relevant web sites. LocusLink is a collaborative effort among NCBI, Human Gene Nomenclature Committee, OMIM, and others. LocusLink currently contains data for a number of species such as human, mouse, rat, zebrafish, nematode, fruit fly, cow, sea urchin, African clawed frog, and HIV-1. Organisms can be searched together or separately.

LocusLink can be accessed via <http://www.ncbi.nih.gov/LocusLink/>

UniGene

UniGene is another NCBI databases where ESTs and full-length mRNA sequences organized into clusters that each represent a unique known or putative gene within the organism from which the sequences were obtained. UniGene clusters are annotated with mapping and expression information when possible (e.g., for human), and include cross-references to other resources. Sequence data can be downloaded by cluster through the UniGene web pages, or the complete data set can be downloaded from the repository/UniGene directory of the FTP site. In addition, UniGene DDD (described below) can be used to show differential expression of genes between cDNA libraries. The organisms represented in UniGene are listed on the UniGene home page.

UniGene can be accessed via <http://www.ncbi.nih.gov/UniGene/>

HomoloGene

HomoloGene is a gene homology tool that compares nucleotide sequences between pairs of organisms in order to identify putative orthologs. Curated orthologs are incorporated from a variety of sources via LocusLink. Organisms represented are listed on the HomoloGene home page.

HomoloGene can be accessed at <http://www.ncbi.nih.gov/HomoloGene/>

OMIM

Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders. OMIM focuses primarily on inherited, or heritable, genetic diseases. It is also considered to be a phenotypic companion to the human genome project. OMIM is based upon the text Mendelian Inheritance in Man, authored and edited by Dr. Victor A. McKusick and a team of science writers and editors at Johns Hopkins University and elsewhere.

OMIM (Online Mendelian Inheritance in Man) is a computerized database version of Victor McKusick's book, Mendelian Inheritance in Man, provided through the National Center for Biotechnology Information. The primary difference between the two resources is that the online version is more current. The online database is updated daily, whereas the book contains all the information that was available online at the time of print. The online version also provides links to a variety of related resources. The print version contains a foreward, preface, and appendices that are not available online.

OMIM is reached at <http://www.ncbi.nlm.nih.gov/omim/>

PubMed

PubMed is a database of citations and abstracts for biomedical literature. These citations are from MEDLINE and additional life science journals. PubMed also includes links to many sites providing full text articles and other related resources. PubMed is accessible through the Entrez search and retrieval system.

PubMed is reached at <http://www.ncbi.nih.gov/entrez/>

Entrez Utilities

Entrez Programming Utilities, also called E-Utilities, are tools that provide access to Entrez data outside of the regular web query interface. They represent a method of making WWW links to Entrez. Each utility performs a specialized retrieval task, and can be used simply by writing a specially formatted URL. For example, EFetch retrieves records in the requested format from a list of one or more primary IDs or from the user's environment. The E-Utilities web page describes the available utilities and links to a brief help document for each one. E-Utilities can be helpful for retrieving search results for future use in another environment.

For more information about E-Utilities, go to

http://www.ncbi.nih.gov/entrez/query/static/eutils_help.html

11. Tables

Table 1: List of the original 11 PGA centers:

- BayGenomics (UCSF)
- Berkeley PGA
- CardioGenomics (Harvard)
- HopGenes (Johns Hopkins)
- InnateImmunity (U. Arizona)
- JAX PGA
- ParaBioSys (Mass. General Hospital)
- PhysGen (U. Wisconsin)
- SeattleSNPs (U. Washington)
- Southwestern (UT Southwestern)
- TREX (TIGR)

Table 2: PGA Topics Under Investigation:

- Animal models and phenotypes
- Clinical and physiological studies
- Databases and software tools
- Expression Profiling
- Mutagenesis
- Proteomics
- SNP and Genotypes
- Comparative Sequence Analysis
- Education Programs

Table 3: Comparison of sample clinical ontologies

	MeSH	Snomed	ICD-9	UMLS
Size	12-25mb	footnote 2	1gb	20gb
Cost	Free	License	Free	License
Usage	Literature	Clinical	Billing	Multipurpose
Ease of use	Easy	?	Easy	Painful
# concepts	22,586	357,000	footnote3	> 1 million
Semantic Relationships	No	Yes	No	Yes

² Size of SNOMED unknown by manuscript author, as that information cannot be found on their website and it was not used at all for this project.

³ Unknown number of concepts in ICD-9 by manuscript author, as that information is not on official website, but it is easily more than the number in MeSH, and likely approximates the number in SNOMED.

Table 4: List of Python scripts and modules from MeshLinker

retrieve.py	Master backend script			backend
dbfunctions.py	Contains reused functions for interfacing with MySQL database			backend
gbpmid.py	Retrieves info from GenBank and PubMed via e-Utils	GenBank accession #	List of PubMed ID #'s	backend
pubmed.py	Retrieves abstract information and MeSH headings via e-Utils	PubMed ID	Author, Title, Journal, Volume, Page, Year, MeSH headings	backend
grabber.py	Retrieves info from Entrez Gene, UniGene, HomoloGene	LocusLink ID	UniGene name, UniGene ID, HomoloGene ID, OMIM ID	backend
meshfill.py	Reads in entire official MeSH XML file (desc2004.xml), which is about 150mb in size. Parses out all synonyms associated with the descriptors, and MeSH ID numbers associated with descriptors. Deposits results into tables Descriptor, MeshId, and Synonym			backend
header.py	Contains HTML for top portion of web page (title and search box)			frontend
search.py	Processes search queries and returns HTML page with results	text query		frontend
meshBrowser.py	Renders webpage, provides search capability, and allows users to browse up and down MeSH tree	MeSH number	Browser page showing category belonging to MeSH number	frontend

Table 5: MeSH Categorizations of Baygenomics Genes

Total number of genes in MeSH Category “C” (Diseases): 928

Disease MeSH Category	Category Number	Number of Genes	Percentage of total
Nervous System Diseases	C10	421	45.3
Immunologic Diseases	C20	350	37.7
Neoplasms	C04	302	32.5
Pathological Conditions, Signs and Symptoms	C23	281	30.3
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16	178	19.2
Disorders of Environmental Origin	C21	102	10.9
Hemic and Lymphatic Diseases	C15	83	8.9
Animal Diseases	C22	65	7
Female Genital Diseases and Pregnancy Complications	C13	64	6.9
Cardiovascular Diseases	C14	63	6.8
Digestive System Diseases	C06	59	6.4
Nutritional and Metabolic Diseases	C18	59	6.4
Urologic and Male Genital Diseases	C12	57	6.1
Endocrine Diseases	C19	53	5.7
Musculoskeletal Diseases	C05	45	4.8
Skin and Connective Tissue Diseases	C17	38	4.1
Respiratory Tract Diseases	C08	27	2.9
Eye Diseases	C11	25	2.7
Otorhinolaryngologic Diseases	C09	16	1.7
Virus Diseases	C02	16	1.7
Stomatognathic Diseases	C07	13	1.4
Bacterial Infections and Mycoses	C01	12	1.3
Parasitic Diseases	C03	6	0.6

Table 6: BayGenomics cell lines with known phenotypes

Cell line	Protein Name	Disease linkages	Gene Accession
RRK003	Hemoxygenase	Linked to acute lung injury	NM_010442, M33203, X56826
RRN335	MIF	Linked to ischemic heart disease	AF204395
Ex318	Fibulin 1	Knockout causes developmental emphysema and Marfan's Syndrome	AK083573, AK08641, NM_01180
TEA176	Integrin Alpha 6	Linked to bullous skin disease, also present in knockout	NM_008397
RRS565	TGFbeta3	Knockout causes cleft palate and hypoplastic lungs	AJ414642

Table 7: List of German Gene Trap Consortium cell lines with known phenotypes

GGTC Line	Gene Symbol	Phenotype
A006B04	Spry4	limb deformation
A20010	novel	pigmentation
W027B02	Neph	podocyte fusion
W036C08	BRC	placenta
W044B06	NCH	lacrima gland
W073D02	DMP1	senescence
Jumonji	JMJ	tube closure
M004D05	STAF	sterility
M016A06	PHD finger	dwarfism
3C7	LTBP4	colon cancer, pulmonary emphysema
W024F10	PKP4	diabetes

12. Figures/Illustrations

Figure 1: Programs for Genomic Applications

Screenshot depicting the homepage for the NHLBI PGA web site. The PGA is a major initiative to advance functional genomic research related to heart, lung, blood, and sleep health and disorders. Goals include developing information, tools, and resources to link genes to biological function on a genomic scale.

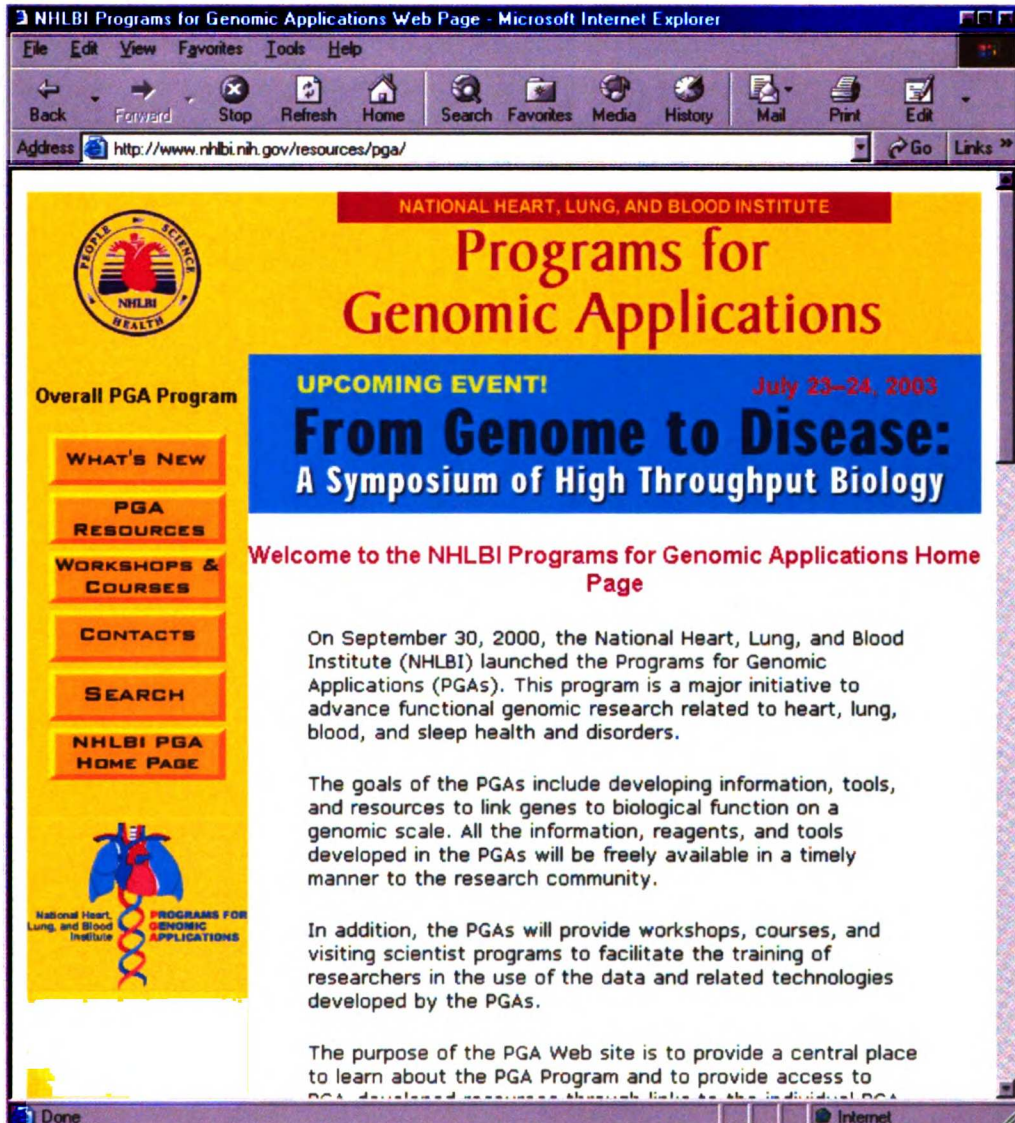


Figure 2: BayGenomics

Screenshot depicting the homepage of the UC San Francisco PGA center, BayGenomics. The major goal of BayGenomics is to identify genes relevant to cardiovascular and pulmonary disease. BayGenomics uses gene-trap vectors to inactivate thousands of genes in mouse embryonic stem (ES) cells for the purpose of generating knockout mice.

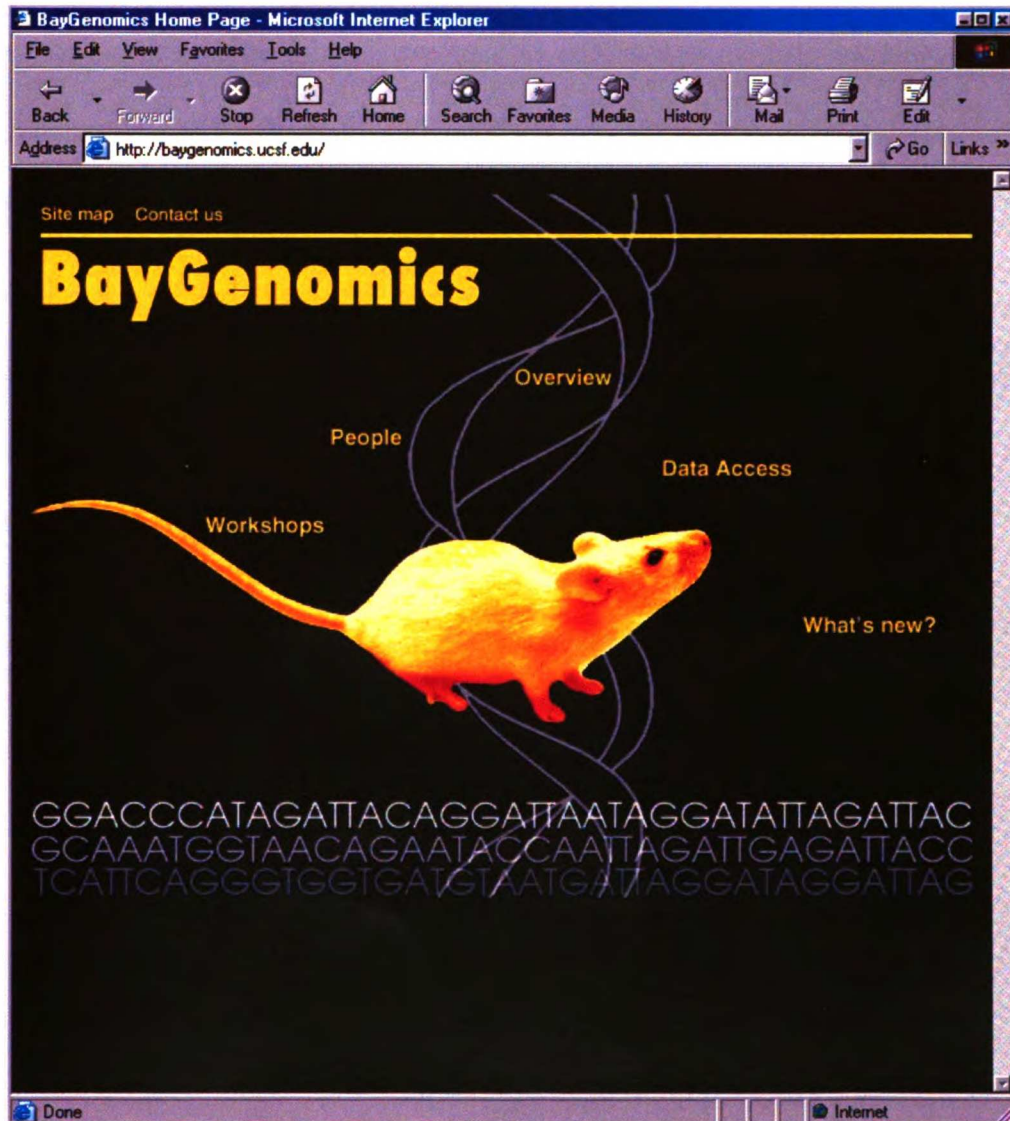
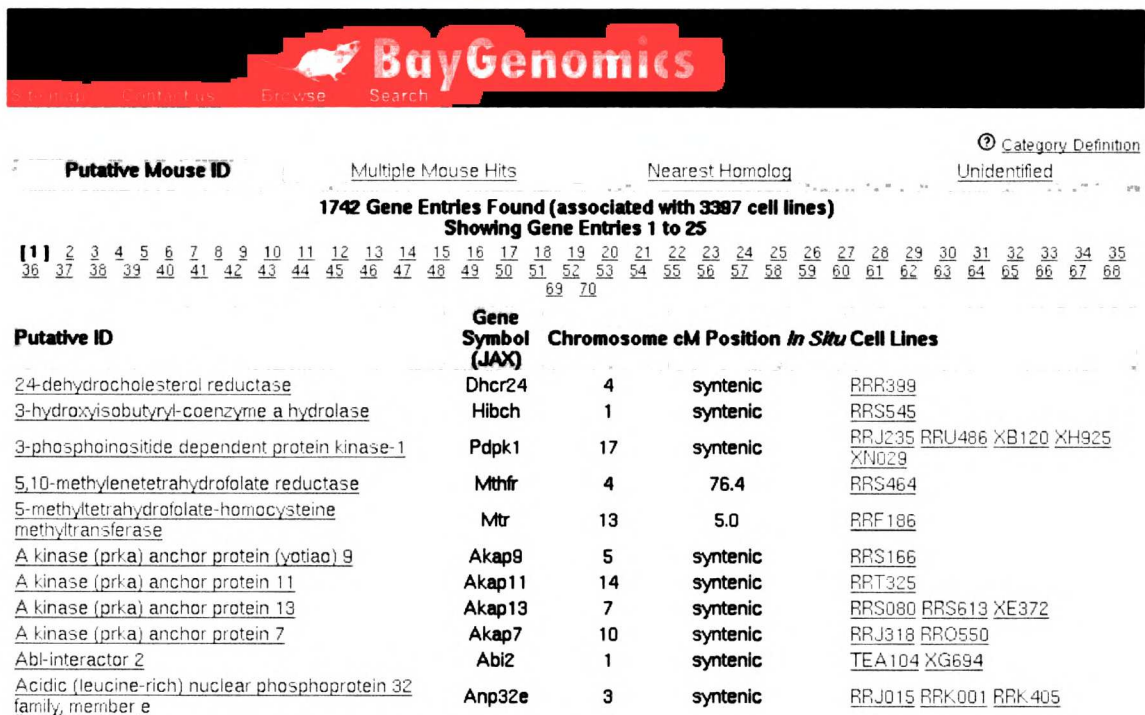


Figure 3: Screenshot of a BayGenomics data access page

By clicking on the “Data Access” link from the main page, and then using the browse function, the user is brought to this page in BayGenomics. Here, one can clearly see the various gene names on the left-most column, and the cell lines associated with that particular gene, on the right-most column.



The screenshot shows the BayGenomics website interface. At the top, there is a navigation bar with the BayGenomics logo and links for Home, Data Access, Browse, and Search. Below the navigation bar, there are several tabs: Putative Mouse ID (selected), Multiple Mouse Hits, Nearest Homolog, and Unidentified. A notification indicates that 1742 gene entries were found, associated with 3387 cell lines, and the first 25 entries are shown.

Putative ID	Gene Symbol (JAX)	Chromosome	cM Position	<i>In Situ</i> Cell Lines
24-dehydrocholesterol reductase	Dhcr24	4	syntenic	RRR399
3-hydroxyisobutyryl-coenzyme a hydrolase	Hibch	1	syntenic	RRS545
3-phosphoinositide dependent protein kinase-1	Pdpk1	17	syntenic	RRJ235 RRU486 XB120 XH925 XN029
5,10-methylenetetrahydrofolate reductase	Mthfr	4	76.4	RRS464
5-methyltetrahydrofolate-homocysteine methyltransferase	Mtr	13	5.0	RRF186
A kinase (prka) anchor protein (yotiao) 9	Akap9	5	syntenic	RRS166
A kinase (prka) anchor protein 11	Akap11	14	syntenic	RRT325
A kinase (prka) anchor protein 13	Akap13	7	syntenic	RRS080 RRS613 XE372
A kinase (prka) anchor protein 7	Akap7	10	syntenic	RRJ318 RRQ550
Abl-interactor 2	Abi2	1	syntenic	TEA104 XG694
Acidic (leucine-rich) nuclear phosphoprotein 32 family, member e	Anp32e	3	syntenic	RRJ015 RRK001 RRK405

Figure 4: Screenshot of a BayGenomics annotation page

By using the browse function or searching capabilities of BayGenomics, the user is presented with an annotation page resembling this one. This particular page shows the annotation information in Baygenomics for cell line RRS464. One can see the links to outside genome localization tools such at UCSC Blat and Ensembl, and also to the usual NCBI databases such as GenBank and LocusLink.

GeneTrap Resource DB Results: RRS464

Cell Line: RRS464 (240 bp)

Genome Mapping: [chr4: 76,400,000](#)

Vector: [pMT2 \(Puro\)](#)

Species: [Mus musculus](#)

Category: [Pituitary Mouse](#)

Identification: [5,10-methylenetetrahydrofolate reductase](#)
[GenBank: U01700.4](#), LocusID: [10233](#) - 100.0 % identity over 100.0 % length of the RRS464 sequence tag

Gene Symbol (JAX): [Mthfr](#)

Chromosome: [4](#)

cM Position: [76.4](#)

Top Mouse dbEST Match: [Mthfr](#) - matched 100.0% identity over 100.0% length of the RRS464 sequence tag [[BLAST](#)]

dbGSS: [Mthfr](#)

Sequence:
>RRS464 (Mthfr; 5,10-methylenetetrahydrofolate reductase)
GGGTTATGTCTTCCAGAAGGCCTACCTCGAATTCTTCACCTCCCGTGAAACTGTGGAGGCGCTTCTGCAGGTGCTGAAG
ACATAACGAGCTGCGGGTCAACTACCAACATCGTGGACGTGAAGGGAGAGAACATCACTAATGCCCTGAGCTGCAGCCCAA
TGCCGTGACGTGGGGCATCTTCCCGGGTCGAGAGATCATCCAGCCTACTGTGGTGGACCCCATCAGCTTCATGTTCTGGA
[Download](#) or [view](#) [ABI trace file](#) used for RRS464.

Post Date: [Oct 10, 2003](#)

Last Updated: [Jun 18, 2004](#) ([History](#))

Availability: This sequence was obtained by 5'RACE PCR from a mouse embryonic stem (ES) cell clone with an insertional mutation from a gene trap vector. [Sequence](#) [files](#) are [available](#) to the scientific community for the purpose of generating a gene knockout mouse.

Copyright 2002 Regents of the University of California. All rights reserved.

Figure 5: Flowchart for data retrieval

Diagram depicting the overall flow of information. Starting with the initial GenBank accession number, the data retrieval is accomplished through two pathways. The first is by obtaining the LocusLink ID, then hitting the UniGene website to obtain the HomoloGene ID, then finding the human gene homolog in order to determine the corresponding OMIM number. The second pathway takes the GenBank accession number, obtains the various PubMed numbers within the corresponding GenBank annotation page, and then downloads all associated PubMed abstracts, eventually obtaining the MeSH headings contained in each abstract. In each case, the obtained data is then stored in the SQL database.

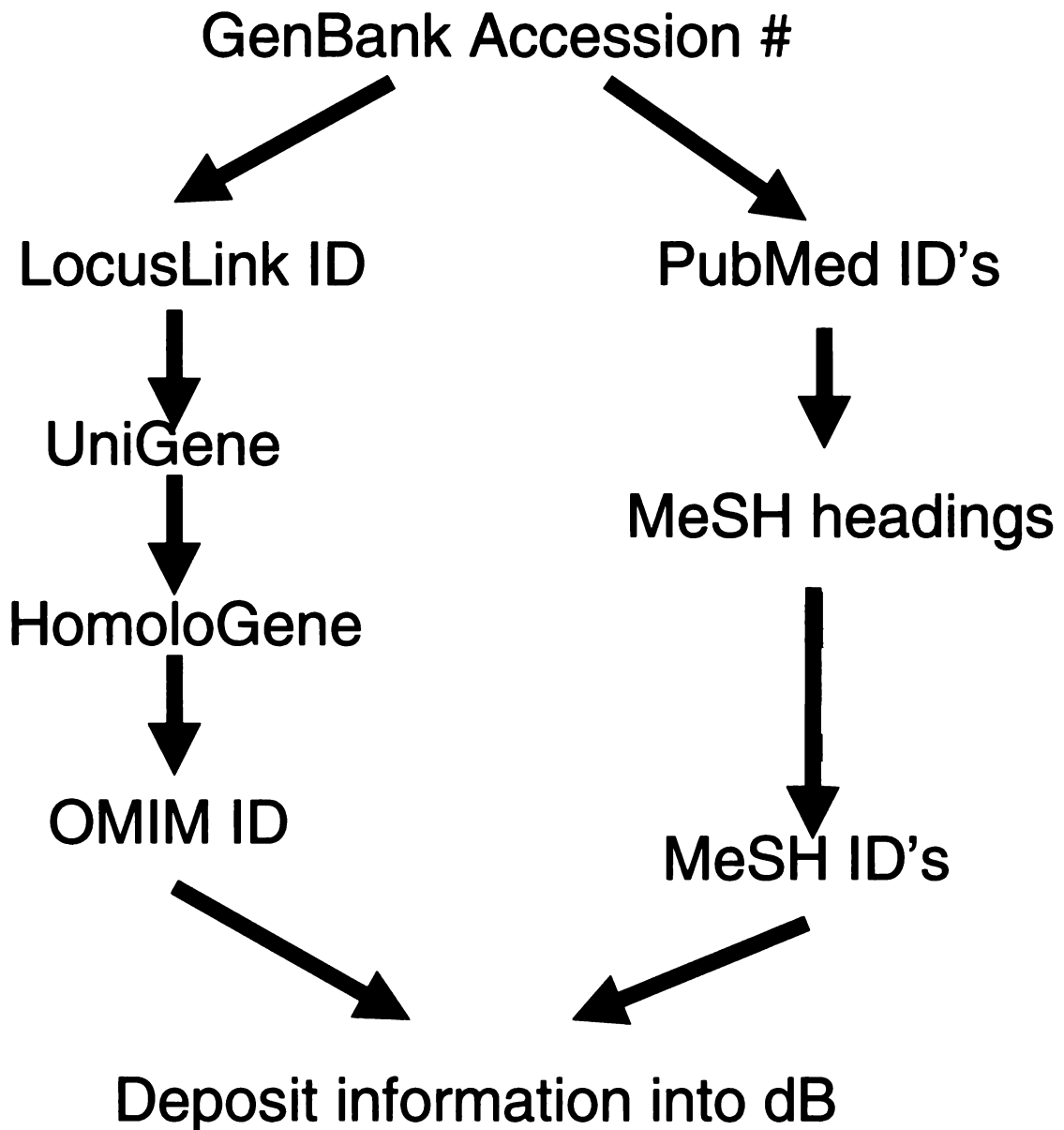


Figure 6: Database schema

Diagram showing the MeshLinker database schema. There are overall 3 main categories, shown in the green, blue, and red boxes. The 3 tables in the green area are taken directly from the production BayGenomics server. The tables in the red area correspond to the information obtained from the first pathway described in Figure 5 (Unigene, HomoloGene, and OMIM). The tables in the blue area correspond to the information obtained in the second described pathway (abstract information and MeSH headings).

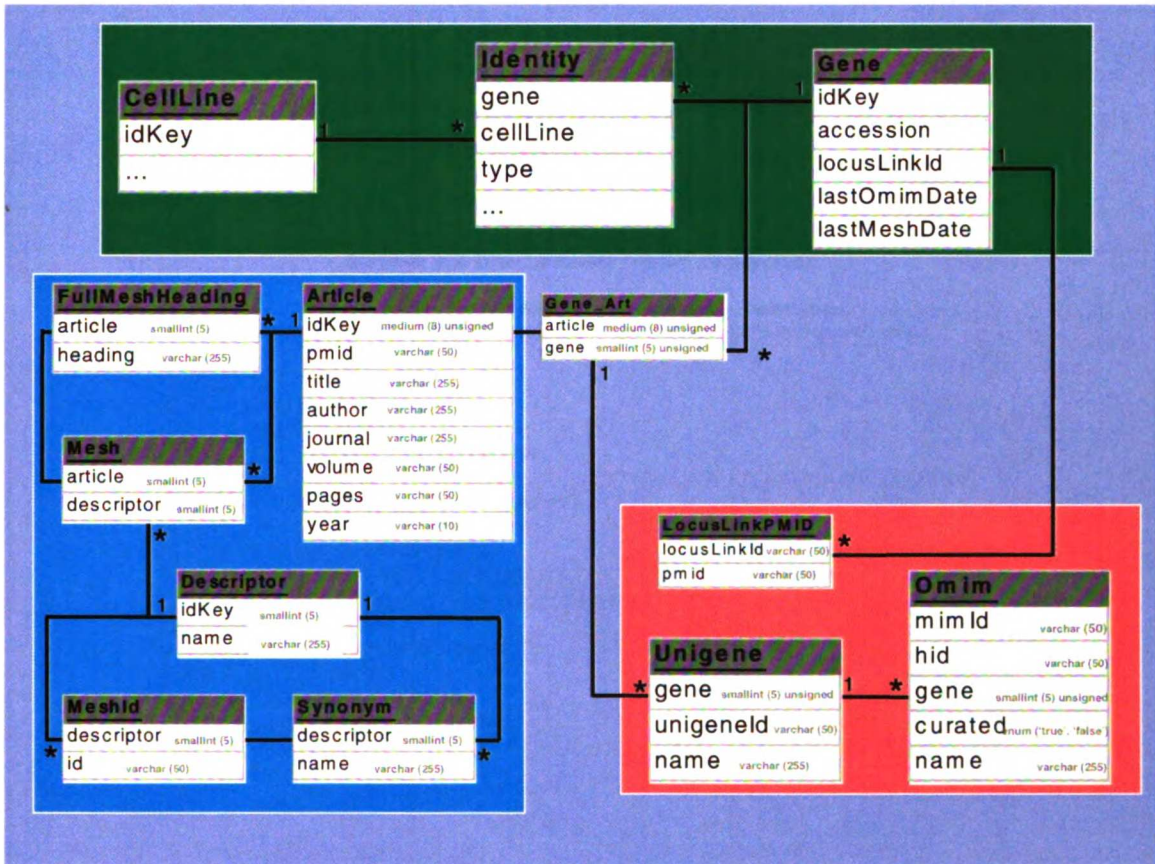


Figure 7: MeshLinker title page

The main page of MeshLinker resembles many common internet search engines – at the top is a search box that one can input queries for certain diseases and gene accession numbers. Beneath that is a clickable browser that one can use to traverse the entire MeSH tree in search of genes. Here, the top level of the MeSH hierarchy is shown.

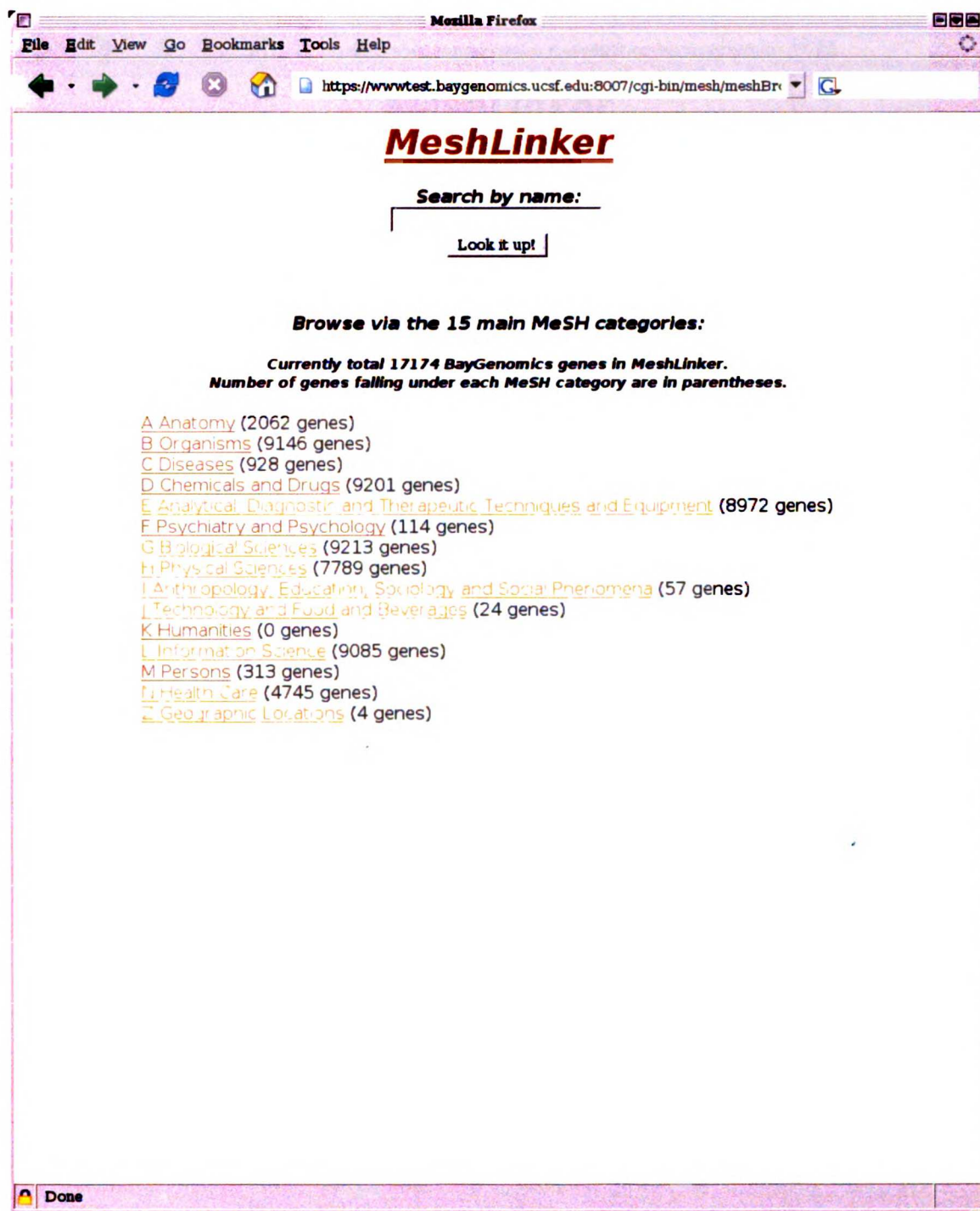


Figure 8: Example browser page displaying MeSH subcategory for “Diseases”

From the first page, clicking on the category “Diseases (C)” will bring the user to the page, showing one step deeper in the MeSH hierarchy.

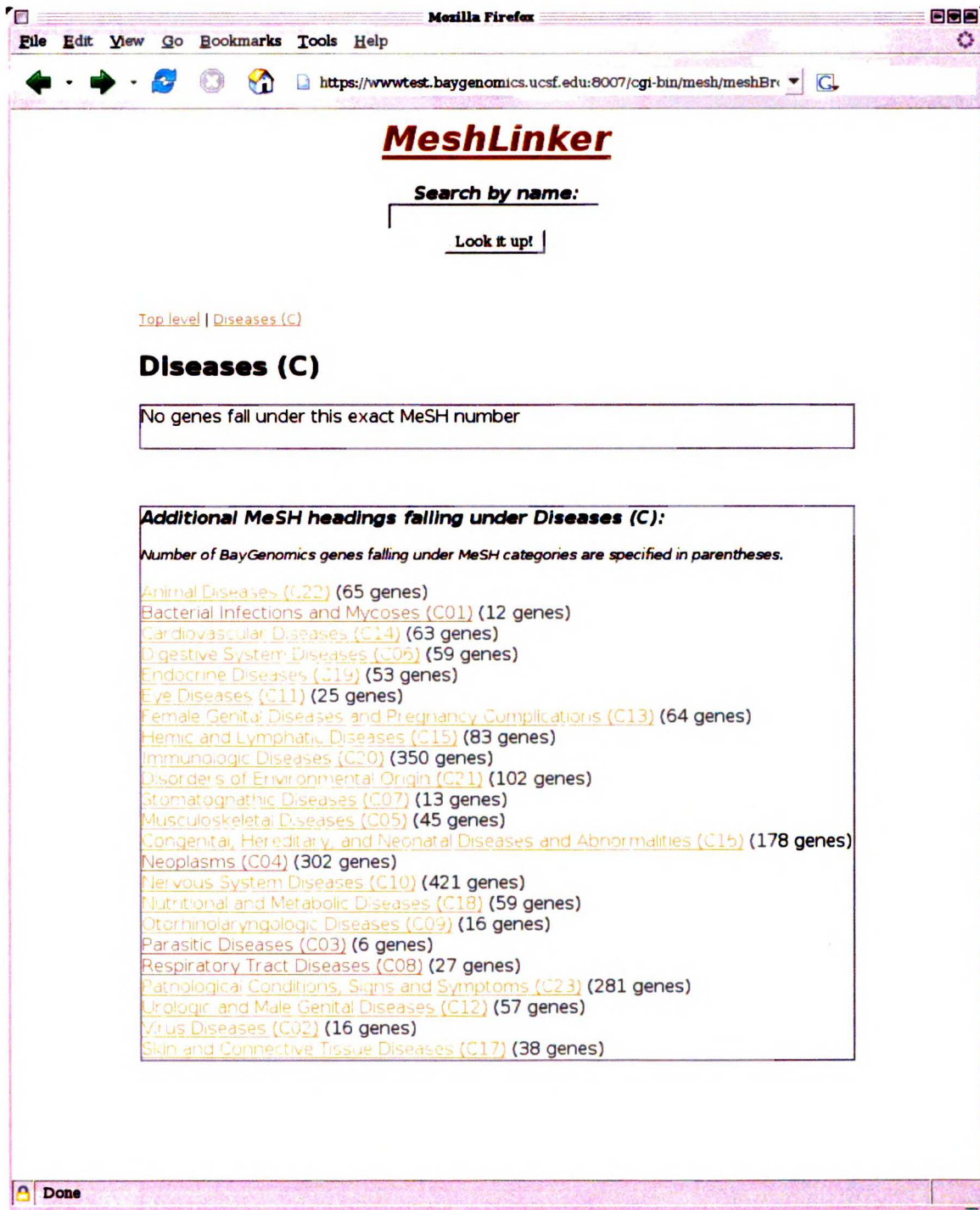


Figure 9: Browser page with gene annotations

Screenshot showing a MeSH category with two genes that have PubMed abstracts containing the exact same MeSH heading name. Dynamic links are generated, leading to outside databases such as Unigene and OMIM, as well as direct links to the associated PubMed abstract.

MeshLinker

Search by name:

[Top level](#) | [Diseases \(C\)](#) | [Respiratory Tract Diseases \(C08\)](#) | [Respiration Disorders \(C08.618\)](#)

Respiratory Insufficiency (C08.618.846)

Genes falling under this exact MeSH number:

[NM_010493](#) Mus musculus intercellular adhesion molecule (Icam1), mRNA.

Cell lines: [M1102 \(Multiple C\)](#), [E285 \(Multiple C\)](#), [R1192 \(Multiple C\)](#)

Unigene Clusters: [M119259](#) Intercellular adhesion molecule (Icam1)

Possible OMIM associations: [147840](#) Malaria, cerebral, susceptibility to.

Pubmed entries linking gene to MeSH number: [Burgin et al](#) Effects of intercellular adhesion molecule 1 (ICAM-1) null mutation on radiation-induced pulmonary fibrosis and respiratory insufficiency in mice. *J Natl Cancer Inst* 94, 733-41 (2002).

[NM_033398](#) Mus musculus phosphatidylserine receptor (Ptgsr), mRNA.

Cell lines: [R1192 \(Multiple C\)](#)

Unigene Clusters: [M112429](#) Phosphatidylserine receptor (Ptgsr)

Possible OMIM associations: None

Pubmed entries linking gene to MeSH number: [Liu et al](#) Phosphatidylserine receptor is required for clearance of apoptotic cells. *Science* 302, 1560-3 (2003).

Additional MeSH headings falling under Respiratory Insufficiency (C08.618.846):

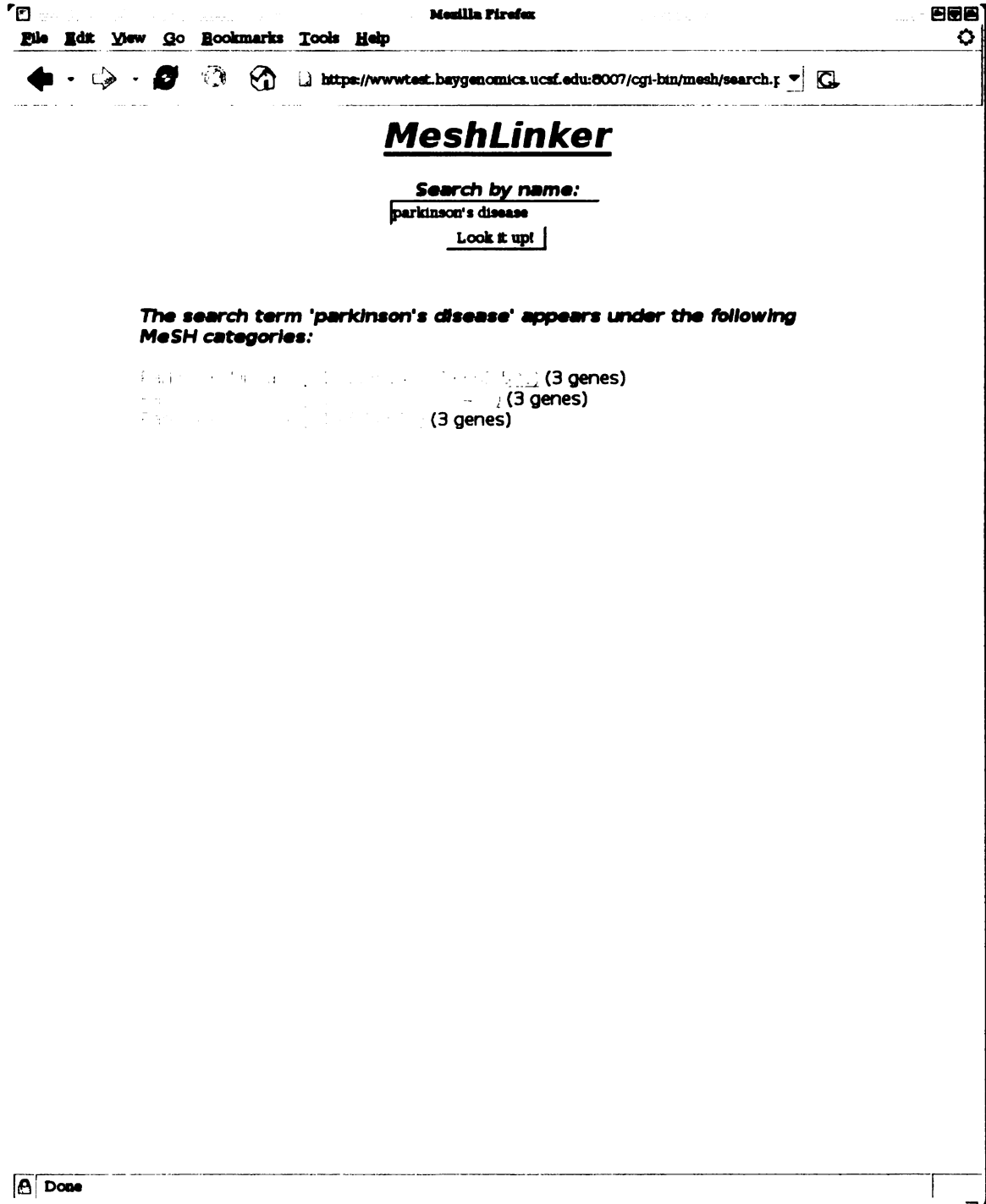
Number of BayGenomics genes falling under MeSH categories are specified in parentheses.

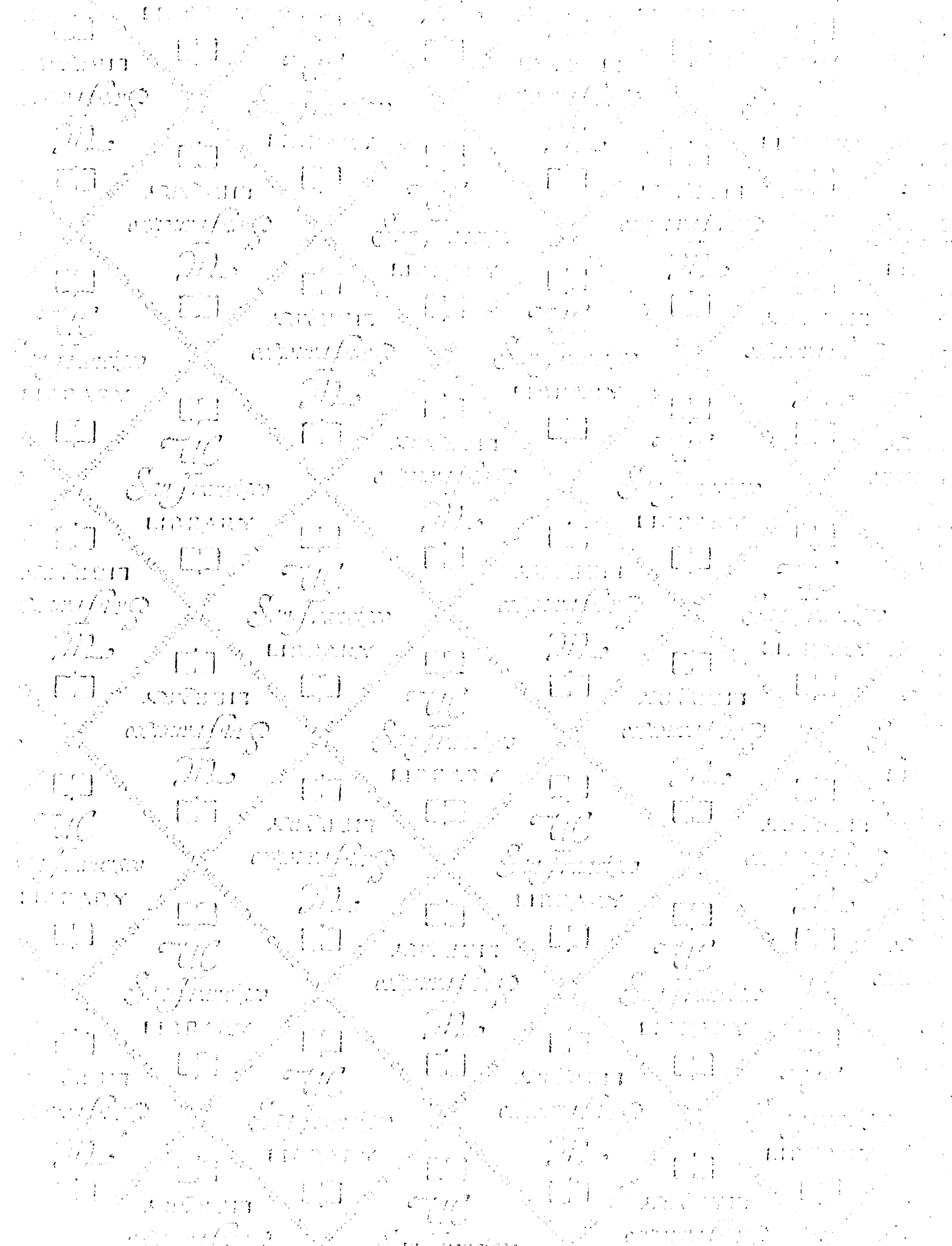
[Respiratory \(C08.618.846.106\)](#)
[Airway Obstruction \(C08.618.846.185\)](#)
[Bronchial hyperreactivity \(C08.618.846.114\)](#)
[Respiratory \(C08.618.846.150\)](#)
[Hypoxia \(C08.618.846.102\)](#)
[Respiratory \(C08.618.846.634\)](#)
[Respiratory \(C08.618.846.112\)](#)

<http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=147840>

Figure 10: Example search results page using query of “parkinson’s disease”

Screenshot showing a search result using the query “parkinson’s disease”. Notice that MeshLinker makes use of the synonyms included with MeSH. The official MeSH term is *Parkinson Disease*, but the application is able to understand the term *Parkinson’s Disease*.





7352209



3 1378 00735 2209

