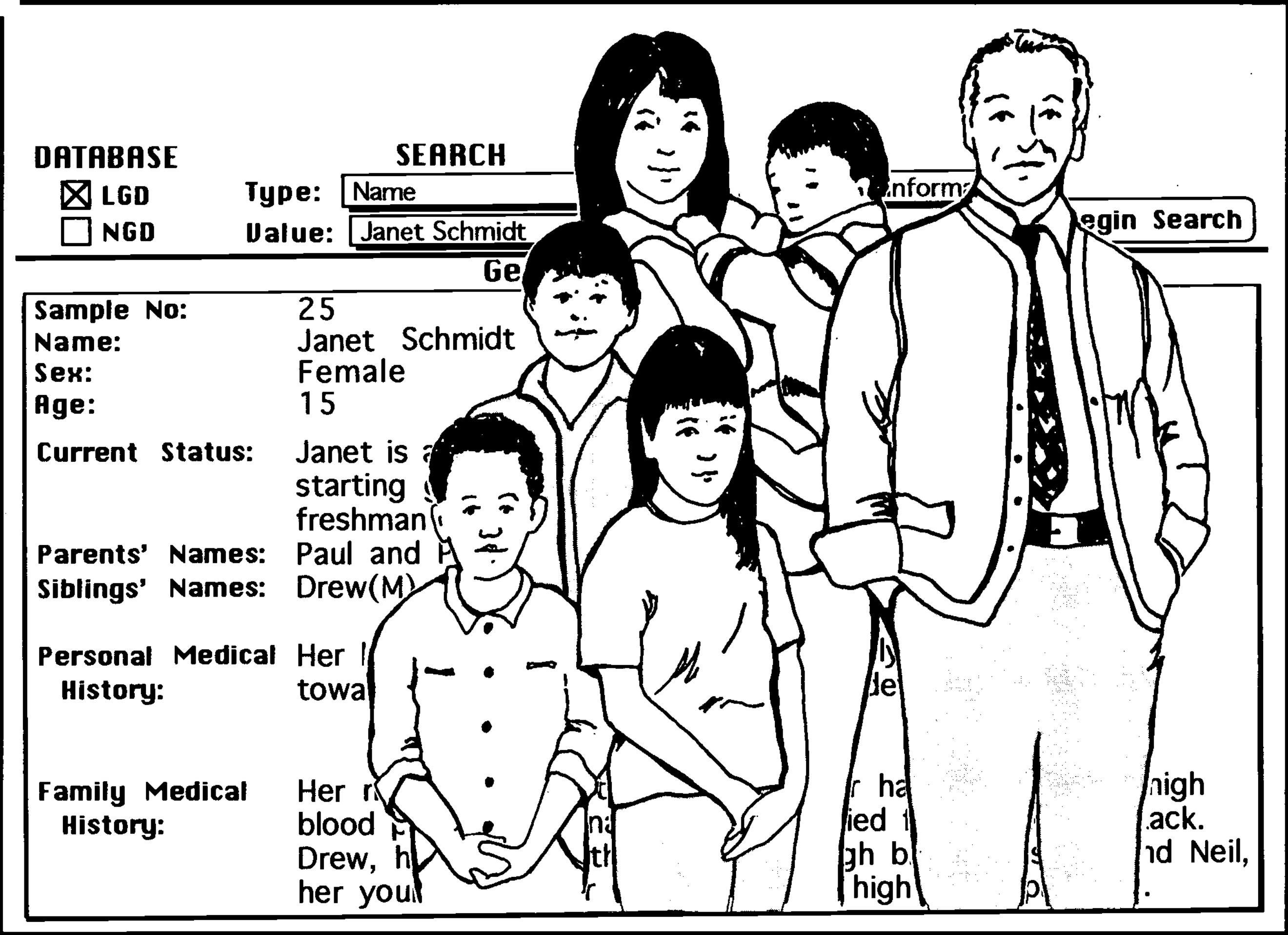
DOCUMENT RESUME

ED 461 499	SE 061 488
AUTHOR	Cutter, Mary Ann G.; Drexler, Edward; Gottesman, Kay S.; Goulding, Philip G.; McCullough, Laurence B.; McInerney, Joseph D.; Micikas, Lynda B.; Mural, Richard J.; Murray, Jeffrey C.; Zola, John
TITLE	The Human Genome Project: Biology, Computers, and Privacy.
INSTITUTION	Biological Sciences Curriculum Study, Colorado Springs.
SPONS AGENCY	Department of Energy, Washington, DC.
PUB DATE	1996-00-00
NOTE	212p.; Distribution of material supported the U.S. Department of Energy and the University of Iowa Genome Center. Accompanying software not available from ERIC.
CONTRACT	DE-FG03-93ER61584
AVAILABLE FROM	Biological Science and Curriculum Study, Colorado Springs, CO 80918-3842. Tel: 719-531-5550; Fax: 719-531-9104; e-mail: info@bscs.org; Web site: http://www.bscs.org. For full text: http://www.bscs.org/pdf/projects/HGN2/HGN-II.pdf.
PUB TYPE	Guides - Classroom - Learner (051) -- Guides - Classroom - Teacher (052)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	*Biology; *Ethics; Evolution; Genetic Engineering; *Genetics; High Schools; *Public Policy; Science Activities; *Science and Society; Science Curriculum; Science Education; Technology
IDENTIFIERS	Biological Sciences Curriculum Study; *Human Genome Project

ABSTRACT

This module, for high school teachers, is the second of two modules about the Human Genome Project (HGP) produced by the Biological Sciences Curriculum Study (BSCS). The first section of this module provides background information for teachers about the structure and objectives of the HGP, aspects of the science and technology that underlie the HGP, information science as it relates to the HGP, and ethics and public policy, especially as they concern information science and the HGP. This section also provides general information on teaching the activities provided in this module, general guidance for discussing values and controversial issues in the classroom, and specific assistance with the supporting software that can be used with this module. The second section of the module includes seven annotated classroom activities plus database software pertaining to genomic registries, explaining the outliers, genetic anticipation, control of information about genes, making public policy, and HGP data and evolutionary biology. (DDR)



DATABASE

☒ LGD

□ NGD

SEARCH

Type: Name

Value: Janet Schmidt

Begin Search

Sample No:	25
Name:	Janet Schmidt
Sex:	Female
Age:	15
Current Status:	Janet is a starting freshman
Parents' Names:	Paul and P
Siblings' Names:	Drew(M)
Personal Medical History:	Her ly towa de
Family Medical History:	Her r r ha high blood ied t ack. Drew, h gh b nd Neil, her you high .

The Human Genome Project:

Biology, Computers, and Privacy

BSCS

The Human Genome Project: Biology, Computers, and Privacy

BSCS
Pikes Peak Research Park
5415 Mark Dabling Blvd.
Colorado Springs, Colorado 80918-3842

Authors

Mary Ann G. Cutter, Ph.D.
University of Colorado, Colorado Springs
Colorado Springs, Colorado

Edward Drexler
Pius XI High School
Milwaukee, Wisconsin

Kay S. Gottesman
Genome Database
Johns Hopkins University
Baltimore, Maryland

Philip G. Goulding
BSCS
Colorado Springs, Colorado

Laurence B. McCullough, Ph.D.
Center for Ethics, Medicine, and Public Issues
Baylor College of Medicine
Houston, Texas

Joseph D. McInerney
BSCS
Colorado Springs, Colorado

Lynda B. Micikas, Ph.D.
BSCS
Colorado Springs, Colorado

Richard J. Mural, Ph.D.
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Jeffrey C. Murray, M.D.
University of Iowa
Hospitals and Clinics
Iowa City, Iowa

John Zola
The New Vista High School
Boulder, Colorado

Development of this material was supported by the United States Department of Energy under grant number DE-FG03-93ER61584. Distribution of the material was supported by the United States Department of Energy and the University of Iowa Genome Center.

Any opinions, findings, and conclusions or recommendations expressed in the publication are those of the authors and do not necessarily reflect the views of the United States Department of Energy or of the University of Iowa Genome Center.

BSCS Administrative Staff

Timothy H. Goldsmith, Ph.D., *Chairman*, Board of Directors
Joseph D. McInerney, *Director*
Larry Satkowiak, *Chief Financial Officer*

Project Advisory Committee

Ken Bingman, Shawnee Mission West High School, Shawnee Mission, Kansas
Kay S. Gottesman, Genome Database, Johns Hopkins University, Baltimore, Maryland
Walter Hartung, Nederland High School, Nederland, Colorado
Claire O. Leonard, M.D., University of Utah, School of Medicine, Salt Lake City, Utah
Richard J. Mural, Ph.D., Oak Ridge National Laboratory, Oak Ridge, Tennessee
Jeffrey C. Murray, M.D., University of Iowa Hospitals and Clinics, Iowa City, Iowa
Philip Reilly, M.D., J.D., Shriver Center, Waltham, Massachusetts
Mark Rothstein, J.D., University of Houston Law Center, Houston, Texas

Project Staff

Joseph D. McInerney, Principal Investigator
Lynda B. Micikas, Ph.D., Project Director
Philip G. Goulding, Technology Specialist
Randy K. Backe, Ph.D., Evaluator
Pamela Van Scotter, Editor
Jan Girard, Artist
Dee Miller, Project Secretary

Cooperating Organizations

American Society of Human Genetics
Council of Regional Networks for Genetic Services
National Society of Genetic Counselors

Field-Test Teachers

Jonathan Bealer, Buena High School, Sierra Vista, Arizona
Patty Cameron, San Dieguito High School, Encinitas, California
William Carbone, Glen Rock Junior-Senior High School, Glen Rock, New Jersey
Diana Doepken, Air Academy High School, USAFA, Colorado
Theresa Estes, Southwest Science/Math Magnet, Kansas City, Missouri
Sally House, Coconut Creek High School, Coconut Creek, Florida
Jessie Klein, Ph.D., Middlesex Community College, Bedford, Massachusetts
Theresa Knapp, Adlai E. Stevenson High School, Lincolnshire, Illinois
Dennis McKinney, Orcas Island High School, Eastsound, Washington
Brian Shmaefsky, Ph.D., Kingwood College, Kingwood, Texas
Gary Stellern, John Muir High School, Pasadena, California
Kelly Weiler, Garfield Heights High School, Garfield Heights, Ohio

Figure Credits

Figure 1: onion root tip - John P. Limback (BIODISC)
Figure 1: metaphase chromosome - Gunther F. Bahr
Computer-generated art - Paige Thomas

Production

Jan Girard, cover design
Angela Greenwalt, typesetting and layout
Joy Rasmussen, production support
Judy Rasmussen, production support
Barbara Resch, proofreading
Amy Short, production support

Software Development

Learning Systems Consultants, Inc., Colorado Springs, Colorado
Bob Emmot
Adam Hughes
Robert Schoolfield
Jeff Thomas

Table of Contents

Annotated Student Activities

Foreword

It now is a commonplace to say that we live in an information age and that electronic management of information will be a central feature of life in the twenty-first century. The Clinton Administration talks of plans to expedite the exchange of information through data superhighways, and already we access multiple databases from our home computers and communicate almost instantly with colleagues around the world through electronic mail.

Increasingly, the electronic management of information is becoming a central, indispensable feature of science as research produces ever more data that must be made accessible to the scientific community at large. The accurate storage and rapid retrieval of scientific data are nowhere more critical than in the Human Genome Project (HGP), whose intent is to map and sequence the 80,000 genes that make up the human genetic endowment. This endowment, the product of 3.5 billion years of evolution, contains approximately three billion nucleotides of DNA, a complete record of which ultimately will reside in electronic databases.

To date, detailed map data and some 460 million nucleotides of sequence data reside in government-supported databases that serve scientists worldwide. Indeed, a substantial portion of the budget for the HGP is devoted to the development of new technologies for information storage and retrieval, just one of many illustrations that the HGP is as much a program of technology development as of basic science.

Although progress in the development of electronic databases has been a boon to the HGP, it also has raised important questions about the personal and institutional applications of such data. As researchers uncover more putative associations between particular segments of the human genome and complex human traits, concern grows about the possible misuse of generalized and personal genomic data in areas such as health care, employment, and insurance.

BSCS has developed this instructional module to introduce high school students to the structure, capabilities, limitations, and implications of genomic databases. We have provided print materials and software to involve students directly in the manipulation of hypothetical genomic data and to help them explore the ethical and policy implications inherent in the growing use of electronic databases.

We are grateful to the United States Department of Energy for its financial support of this module, the second such BSCS program the department has supported. As with the first module, the education committees of the American Society of Human Genetics, the National Society of Genetic Counselors, and the Council of Regional Networks for Genetic Services have provided excellent reviews of these materials, and we thank the individuals involved for their assistance. Colleagues at the Genome Database, at Johns Hopkins University, also provided extensive feedback on the program's software, and we are most grateful for their help. The University of Iowa Genome Center provided generous support for distribution of this module. Finally, we thank the teachers and students across the country who field tested these materials for us and provided candid feedback to help ensure that this program is useful for teachers and students nationwide. We hope that all who use these materials find in them the same level of challenge, stimulation, and enjoyment that we found in developing them.

Joseph D. McInerney
Director, BSCS
Pikes Peak Research Park
5415 Mark Dabling Blvd.
Colorado Springs, Colorado

Timothy H. Goldsmith, Ph.D.
Chairman, BSCS Board of Directors
Yale University
New Haven, Connecticut

Module-at-a-Glance

This is the second module that BSCS has produced about the Human Genome Project (HGP). The first, *Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy*, produced in cooperation with the American Medical Association and funded under a grant from the Department of Energy, was distributed free of charge in October 1992 to some 48,000 high school biology teachers in the United States. This second module, funded by the Department of Energy and distributed with support from the Department of Energy and the University of Iowa Genome Center, also will be made available free of charge to interested teachers.

Although both modules deal with the science, ethics, and public policy of the HGP, they are focused quite differently and have been designed to be used independently. That is, students and teachers do not have to be familiar with the content of the first module in order to use the second. The following table compares the two modules:

Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy	***The Human Genome Project: Biology, Computers, and Privacy***
• Review of major mapping and sequencing techniques underlying genomic research; the development of associated technologies.	• Overview of mapping techniques; expanded discussion of the relationship between genetic and physical mapping.
• Human genetic variation; expected results from the HGP; limits of and opportunities for HGP-related research.	• Informatics of the HGP; descriptions of major genomic databases; the role of databases in scientific research.
• Ethical and public-policy issues raised by genomic research.	• Ethical and public-policy issues related to research databases and genomic registries.
• Four classroom activities focused on scientific, ethical, and public-policy issues related to the HGP.	• Seven classroom activities plus database software focused on the informatics of the HGP and on the related ethical and public-policy issues.

Organization of *The Human Genome Project: Biology, Computers, and Privacy*. The module is divided into two sections. The first section provides background information for the teacher about the structure and objectives of the HGP, aspects of the science and technology that underlie the HGP, informatics as it relates to the HGP, and ethics and public policy, especially as they concern informatics and the HGP. We do not expect teachers or students to master this background material. Rather, we have provided it to make teachers more comfortable with the content of the activities. We do *not* intend that

- □ teachers will use the support material as the basis for a series of lectures;
- □ students will read the support material (although some may wish to); or
- □ teachers will convey all of the information contained in this material to their students in the context of the activities.

The first section of the module also provides general information about teaching the activities (for example, information about scheduling and preparation), general guidance for discussing values and controversial issues in the classroom, and specific assistance with the supporting software. More detailed information on these topics is included in the teacher's annotations for each activity.

The second section of the module includes seven annotated classroom activities, which are designed to be used in sequence (see the following table). The introductory activity and Activities 1-3 require students to interact repeatedly with the software that BSCS has developed for this module and together these activities form a conceptual unit that introduces students to the structure and use of genomic databases. In contrast, Activities 4 and 5 focus on the ethical and public-policy questions that the existence of such databases raises. Although Activities 4 and 5 do not require use of the related software, your students' understanding and appreciation of these ethical and public-policy issues will be enhanced significantly if they complete the preceding activities. Finally, the extension activity involves students again with the software as they consider the use of sequence data to infer genetic relationships among members of the same species and evolutionary relationships among different species.

Overview of the Activities

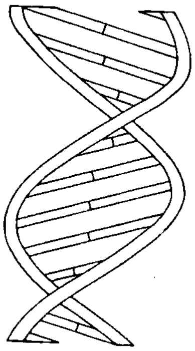
Activity Title	**Description**
Introductory Activity: *The HGP and Electronic Databases*	Students compare the results of a manual search of sequence data with the results of a computer search of the same data and discover the usefulness of electronic storage in handling large amounts of genomic information.
Activity 1: *Genetic Registries*	Students assume the identities of specific individuals in a set of three fictitious families and conduct a series of simple searches to locate information about themselves and their families in a model genetic registry. Students use the data that they retrieve to construct pedigrees for their extended families. In preparation for Activity 4, students also decide whether to authorize further testing of their fictitious person to determine his or her status with respect to the genes for cystic fibrosis, sickle cell disease, and familial hypertrophic cardiomyopathy.
Activity 2: *Explaining the Outliers*	Students compare the pedigrees that they constructed in Activity 1 with those stored in the computer and discover that the model registry contains data that appear to be contradictory. Using both the genetic registry and a model research database as sources of information, the students work together to develop and test a set of hypotheses about the reasons for these contradictions. The searches illustrate the differences between research databases and registries and emphasize the fact that databases can contain errors of various types.
Activity 3: *Genetic Anticipation*	Students investigate the case of a young woman struggling to understand the implications of her genotype for fragile X syndrome. The activity introduces students to the molecular basis of genetic anticipation, first, as a disease mechanism that was discovered as a result of HGP-related research, and, second, as a way to consider some of the potential problems associated with the rapid rate of increase in knowledge about the human genome. Activity 3 also illustrates the role that genomic databases play in the dissemination of new scientific findings and raises questions about the privacy, interpretation, and use of genetic information.
Activity 4: *Who Should Control Information about My Genes?*	Students use the skills of ethical reasoning to analyze a series of options for handling the additional genetic information supposedly gathered about their fictitious families as a result of the voluntary testing initiated in Activity 1. These options range from no disclosure, to private disclosure only to the specific individuals involved, to full, public disclosure by entering the data into the model genetic registry. The activity focuses on different ways to balance the autonomy-based concerns of privacy and confidentiality with the consequence-based concerns of individual and societal health and well-being. The activity also sets the stage for Activity 5, in which students consider the public-policy implications of registry databases.

Overview of the Activities (continued)

Activity Title	Description
Activity 5: *Making Public Policy*	Students compare the implications of three different policy options with respect to genetic registries. The discussion requires students to build on their understanding of the ethical issues involved and also to draw on the scientific and technical knowledge they acquired during the preceding activities. By the end of Activity 5, students should understand that public policy, when done carefully, is a powerful form of "preventive ethics" because effective public policy anticipates and addresses likely as well as unlikely ethical concerns.
Extension Activity: *HGP Data and Evolutionary Biology*	Students use sequence data to infer genetic relationships among members of the same species and evolutionary relationships among different species. In Part A, students use DNA sequence data to propose and test the hypothesis that a set of nine skeletons found in a shallow grave in Russia in 1991 includes the remains of Tsarina Alexandra and three of the Romanov children, assassinated during the Bolshevik revolution in 1917. In Part B, students extend their understanding of the importance of sequence analysis by using both DNA sequence data and amino acid sequence data as the bases for constructing two simple phylogenetic trees.

Masters (BLMs) for the teacher support materials, including worksheets and overhead transparencies, begin on p. TS-1. Masters for the student materials, including the text and worksheets for the activities, begin on p. S-1. The copyright for these materials allows you to make unlimited copies of the BLMs and the related software for classroom use.

An evaluation form begins on p. E-1. Please take a few minutes to complete and return this form after you have used the module.



Section I What Is the Human Genome Project?

The Human Genome Project (HGP) is a large, internationally coordinated effort in biological research directed at creating a detailed map of human DNA. One way to understand the project is to see it as a natural culmination of the revolution in molecular genetics that began in the 1970s and that gave rise to the recombinant-DNA technologies that are so critical to its work. Yet another way to understand the project is to see it as the beginning of a new revolution, a revolution in our way of asking and answering many of the questions that we continue to have about human evolution, development, variation, behavior, and the complex interactions between genes and environment that make us what we are.

Simply mapping and sequencing the entire genome will not give us answers to all of our questions. For example, the HGP will not, in itself, explain the functions of all of the genes that make up a human being, the ways in which these genes interact, nor all of their differences from one of us to the next. It will, however, produce a powerful set of research tools—the human DNA sequence, as well as the sequences of the genomes of several other organisms—that scientists will use to help answer these questions. HGP-derived data will fuel the work of scientists for years to come, and analyzing the data for answers to questions of biological and medical interest will have enormous impact, not only within science, but in many other areas of life as well.

The Objectives of the Human Genome Project

By the early 1980s, advances in several fields had made it possible to conceive of mapping and sequencing the complete human genome. Key among these advances were the increasing ability of scientists to isolate and manipulate specific DNA fragments and the construction of small parts of the human map by several different research groups.

The original proposals for the project focused on plans to sequence the entire genome, that is, to determine the exact order of the estimated three *billion* As, Cs, Gs, and Ts that are strung together to make up the DNA in the human chromosomes. Critics of this approach, however, pointed out that technologies did not yet exist to allow scientists to sequence the three billion nucleotides in the human genome in a rapid and cost-effective way. They also noted that even if we knew the complete sequence, we still would not be sure how to identify the genes (*coding sequences*) among all of the other sequences in the DNA (*noncoding sequences*), nor how to distinguish portions of gene sequences that code for protein (*exons*) from those that do not (*introns*).

Many scientists argued that, until costs could be reduced and the speed could be increased, sequencing a huge amount of unidentified DNA, most of which likely would be noncoding, should have low priority. On the other hand, most scientists involved in early discussions of the HGP agreed that detailed maps of each human chromosome and of the chromosomes of several other well-studied organisms would be extremely useful. Such maps would help biologists begin to answer questions of current interest and also would provide the information required to interpret the full human sequence once it became available.

Therefore, the *primary* goal of the HGP was to develop detailed genetic and physical maps of the human genome as well as of the genomes of *Escherichia coli* (a bacterium), *Saccharomyces cerevisiae* (a yeast), *Caenorhabditis elegans* (a roundworm), *Drosophila melanogaster* (a fruit fly), *Mus musculus* (the laboratory mouse), and *Arabidopsis thaliana* (a rapidly growing plant that has a small genome). At the most basic level, these maps

Genetic Map:	Complete a fully connected human genetic map with markers located an average of two to five centimorgans apart. [A centimorgan is a unit of measurement used on genetic maps. Two genes are located one centimorgan apart if they are separated one percent of the time by crossing-over during meiosis.]
Physical Map:	Complete STS—sequence-tagged site—maps of all human chromosomes, with markers spaced approximately 100,000 bases apart. [An STS is a short segment of a chromosomal DNA molecule whose sequence has been determined and is known to be unique. A set of STSs identified on a chromosomal DNA molecule helps scientists integrate genetic-linkage and physical maps of the chromosome.]
DNA Sequencing:	Develop efficient approaches to large-scale sequencing of DNA, building to a collective capacity of 50 Mb per year. [50 Mb is equal to 50 million bases.]
Model Organisms:	Finish sequencing the *E. coli* and *S. cerevisiae* genomes. Continue sequencing the *C. elegans* and *D. melanogaster* genomes with the goal of bringing them near completion by the end of 1998. Sequence selected portions of *M. musculus* DNA side-by-side with corresponding human DNA in areas that seem to be particularly interesting or important.
Informatics:	Continue to create, develop, and operate databases to provide easy access to up-to-date mapping and sequencing information. Develop effective mechanisms for data exchange and for effective searching among databases. Continue to develop algorithms for comparing and interpreting genomic information.
Ethical, Legal, and Social Implications:	Continue to develop programs to identify and define the ethical, legal, and public-policy issues associated with genomic information. Develop policy options regarding genetic testing and foster greater public acceptance of human genetic variation.

[From: Collins, F. & Galas, D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science* 262 (5130):43-46.]

Table 1 Selected goals for the HGP (1993-1998). In 1990, the National Institutes of Health (NIH) and the U.S. Department of Energy (DOE), the federal agencies primarily responsible for organizing and funding the HGP in the United States, submitted a joint proposal to Congress that outlined specific goals for the first five years of what was expected to be a 15-year effort. Since the project was implemented (1 October 1990), scientists have made significant progress toward achieving its goals. In October 1993, Francis Collins, director of the National Center for Human Genome Research (NIH), and David Galas, formerly associate director of the Office of Health and Environmental Research (DOE), reported that the project was not only on schedule, but in some cases, ahead of schedule. As a result, representatives from NIH and DOE drafted a new set of extended goals to cover the period 1 October 1993 to 30 September 1998. Some of these goals are summarized above.

would assign genes and other markers to particular chromosomes, would establish the order of these genes and markers along the chromosomes, and would provide information about the distances between them. Today, most scientists believe that the work of building these maps is essentially complete.

A second goal of the HGP is to determine the complete base sequence of each of these genomes. This work has now started in earnest; in fact, scientists already have determined the sequence of the 16,569 bases in human mitochondrial DNA. In October 1995, a large team of researchers reported the complete genetic sequence—580,070 bases—of the bacterium *Mycoplasma genitalium*. In March 1996, another team reported the complete sequence —12.5 million bases—of the yeast *Saccharomyces cerevisiae*. (This particular work was the result of an international, collaborative effort.) With recent improvements in automated sequencing, the goal of determining the complete human sequence is now clearly in view.

Even when the human sequence is completed, however, the work of understanding the human genetic legacy will not end. Although it is true that the completed sequence will represent a human genome at its greatest level of detail, this sequence will be interpretable only as it is viewed against the backdrop of the DNA sequence variation that is responsible for both normal and abnormal genetic variation among us.

WHAT WILL WE LEARN FROM THE HUMAN GENOME PROJECT?

It is difficult to predict the impact that the HGP will have on the growth and practice of biology, on clinical medicine, on the science of informatics, and on technology. Scientists already have some sense of its possible results, but they expect that many surprises still await us.

Insights into basic biology. New data generated in conjunction with the HGP will raise a number of interesting biological questions and help scientists address them. An obvious benefit of the work is the discovery of new genes. As laboratory groups identify, map, and sequence these genes, scientists build a foundation for studying gene function and regulation, and, possibly, for discovering and investigating biological phenomena not previously recognized. For example, a recent result to come from work associated with the HGP is the discovery of the molecular mechanism that underlies the phenomenon that scientists call *genetic anticipation*. The term "anticipation" refers to a long-standing observation that the severity of some genetic disorders can increase with each succeeding generation or that the age of onset for the same disorders can decline in succeeding generations. Recent HGP-derived data indicate that genetic anticipation is related to the tendency of certain portions of the DNA to expand beyond their normal length. The discovery of this new and unexpected mechanism for genetic disease supports the argument that work associated with the HGP will not only provide new information about inherited diseases, but also may help us think in new ways about the fundamental mechanisms of genetics. Activity 3 in this module introduces students to the phenomenon of genetic anticipation.

As the HGP progresses, scientists will become better able to recognize genes within DNA sequences and also better able to predict the functions of the proteins specified by the DNA. Scientists estimate that there is now about a 50 percent probability that a newly isolated gene will be related to a gene we already know. As we accumulate more sequence data, cross-relationships among genes will become more evident, which will allow us more accurately to propose functions for new genes and more easily to devise experiments to test these hypotheses.

It is likely that many new insights also will emerge from the close study of the large sections of DNA that have not yet been assigned a function. This noncoding DNA, estimated to represent about 95 percent of the human genome, is sometimes inappropriately called "junk" DNA. Yet, it may contain nucleotide sequences that make all the rest of the DNA work. Currently, the HGP is focusing its sequencing efforts on DNA that is likely to be part of coding regions. Some scientists have argued, however, that it is not clear whether, in the long term, it will be more efficient to determine which parts of the genome are informative and sequence them, or to sequence the entire genome and then find the informative parts.

Scientists expect that their increasing ability to compare the human genome directly with the genomes of other organisms will yield important information as well, both about the locations and functions of genes, and also about evolution, including human evolution. This is why mapping and sequencing the genomes of nonhuman organisms are such important elements of the HGP. The gene for Duchenne muscular dystrophy, for example, codes for a protein that is similar to several proteins that already have been studied in other organisms. And one of the neurofibromatosis genes codes for a protein whose mechanism of action is understood in part from experiments conducted in a number of animal systems. Thus, learning about the genes of an animal such as a mouse or a roundworm can help scientists develop a deeper understanding of similar genes in humans.

Likewise, the more information that we have available about the human genome and the genomes of other organisms, the better we will understand the evolutionary processes that gave rise to the human species. The human genome is an historical product of 3.5 billion years of organic evolution. As species evolve, some parts, including genes, are conserved (that is, they continue to exist). This suggests that a major consequence of understanding the human genome will be an improved understanding of evolution in general, and of human evolution in particular. The extension activity in this module introduces students to the use of sequence data to study evolutionary relationships.

Improved prediction, diagnosis, and treatment of genetic disorders. Access to detailed map and sequence data also will affect the practice of clinical medicine. Finding the genes associated with inherited diseases is a key concern of the HGP. As investigators isolate and sequence more disease-related genes, they are able to ask increasingly sophisticated questions about the structures of these genes and about the structures and functions of their protein products. Although isolating and sequencing a disease-related gene does not assure that scientists will be able to develop a cure, it can speed the process significantly.

Geneticists, of course, had isolated (cloned) a number of genes that are responsible for genetic disorders even before the formal organization of the HGP. Well-publicized successes include the cloning of genes responsible for Duchenne muscular dystrophy (DMD), retinoblastoma (RB1), and cystic fibrosis (CF). The HGP, however, is allowing researchers to search for disease-related genes in a more systematic manner than in the past. Instead of hunting for specific genes on a one-by-one basis (the search for the CF gene alone required eight years and more than $100 million), biologists are moving toward a coordinated effort to find *all* human genes.

The HGP also will lead to the development of rapid, increasingly inexpensive, and more accurate techniques to diagnose genetic disorders in individuals who exhibit certain symptoms. Before geneticists cloned the DMD gene, confirming a diagnosis required expensive and uncomfortable tests, and the tests were inadequate to detect carriers. Now, with only a blood sample, geneticists can detect most mutations associated with DMD very rapidly. Current DNA-based tests for DMD not only clarify the diagnosis quickly, but also enable geneticists to detect carriers within the same family.

Genomic information sometimes can indicate the likelihood that an individual will get a certain disease in the future. For example, individuals who carry the gene responsible for Huntington disease (HD) almost always will experience symptoms of the condition, although geneticists cannot predict accurately the time of onset. Genomic information also can help geneticists identify those individuals who may have an increased susceptibility to disorders, such as heart disease, cancer, and diabetes, that result from complex interactions between genes and the environment. Although geneticists cannot guarantee that symptoms will occur, they can alert such individuals that their risk is greater than that for others in the general population.

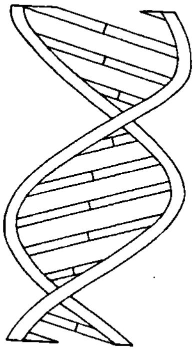
The growing availability of genetic tests that can be used to detect carriers, in prenatal diagnosis, and for presymptomatic identification raises a number of difficult ethical and public-policy issues about the collection and use of such data. It probably is inevitable that our increasing ability to recognize genes that predispose to disease will alter the practice of medicine significantly. It may become routine, for example, to take DNA from newborns and analyze it for allelic forms that can predispose the child to a variety of diseases. Hospitals already test infants born in the United States for phenylketonuria (PKU), a genetic condition that can lead to mental retardation unless the child is placed on a special diet. Thus, we may see medicine slowly moving from a reactive mode (curing people after they are sick) toward a more preventive mode (prescribing therapeutic treatments that can keep people well). Such a shift in emphasis will bring its own set of controversies. Section III of the background information for teachers and Activities 4 and 5 address some of these issues.

Effects on technology. The HGP already has spurred the development of a number of advanced technologies that have had considerable impact outside of the HGP. For example, techniques that allow the detection of minute concentrations of chemicals in liquid environments not only allow those molecules to be identified, but also can provide information about their physical and chemical surroundings. The sensitivity requirement for DNA sequencing—the accurate identification of single molecules—is challenging scientists to push the limits of detection to ever lower chemical concentrations. It may be that one day, investigators will be able to use single molecules as probes to explore biological processes and structures at levels of resolution that are inaccessible by other means.

As in the case of other large, coordinated government efforts, the HGP also will spawn new opportunities for business and industry. These may include opportunities for companies that design and produce biological instrumentation, companies that specialize in robotics and automation, and commercial laboratories and research organizations that work in the areas of DNA diagnostics and biocomputing.

Effects on the practice of science. Finally, many scientists expect the HGP to change the very nature of biological research. We see its impact already in changes in the relationship between molecular biology and the rest of biology. It is becoming increasingly clear that molecular techniques are powerful ways to study many questions in biology, from questions about development to those about evolution. As it becomes easier to ask and answer questions at the level of the DNA, more biologists are becoming, in effect, molecular biologists. It is important to remember, however, that knowledge from molecular biology ultimately must be considered in the context of whole organisms if we are to develop comprehensive insights into biological systems.

Likewise, the HGP is requiring biologists to cooperate and to share information and expertise at levels and in ways that traditionally have not been required. Rather than working independently, investigators are learning to work in loosely organized but highly collaborative groups, often linked by sophisticated computers. The demand for technological development also is intensifying the pressure for biologists to work closely with chemists, physicists, mathematicians, and computer scientists. This trend is affecting the training of new biologists, in effect creating an enormous need for young scientists trained across traditional discipline boundaries. Such training and collaboration may prove to be essential to the progress of biology in the next century, especially as scientists turn their attention to the analysis of other large and complex systems.



Section II The Science and Informatics of the Human Genome Project

It probably is safe to say that what we know about genes, especially human genes, is far less than what remains for us to learn. We do not even know for sure how many human genes there are. Although scientists estimate the number at between 50,000 and 80,000, by January 1996, only some 6,000 human genes had been identified. Of these, only about 4,000 had been assigned to a particular chromosome (some also had been assigned to more specific map positions), and only about 2,000 had been sequenced.[1]

Human genes are located on chromosomes; each chromosome is a single DNA molecule complexed with various proteins. Human cells that have nuclei (except for eggs and sperm) normally contain 22 pairs of autosomal chromosomes, plus 2 sex chromosomes (XX or XY), and many mitochondria, each with a small, circular chromosome. The total genome for any individual includes all of these components. For the purposes of the HGP, however, genome means one each of the different chromosomes—22 autosomes, plus X, plus Y, plus a mitochondrial chromosome. Figure 1 illustrates this human genetic material at increasing levels of detail.

A question that often is asked about the HGP has to do with the source of the DNA that is being mapped and sequenced. The enormous variation among members of the human population is part of our daily experience. It allows us, for example, to distinguish one person from the next, and we recognize that this *phenotypic* variability reflects important underlying *genotypic* differences. Whose genome, then, are scientists studying?

The answer is that no one person's genome is being mapped and sequenced. Although the genomes of two individuals differ in details, the *basic* genetic map for each of us is the same. For example, the major "housekeeping" genes are in the same location for almost all individuals, and functionally important DNA tends to be conserved among humans (that is, most sequence differences tend to occur outside coding regions). Therefore, *mapping* the human genome requires that we map only one of each pair of autosomes, plus the X, Y, and mitochondrial chromosomes, and these chromosomes can come from virtually any person. Likewise, the complete genomic *sequence* will be a composite of sequences derived from many individuals. Although the technologies that result from the HGP eventually may allow researchers to sequence the total genome of a single individual, it is not clear at this point whether such an effort ever would be undertaken.

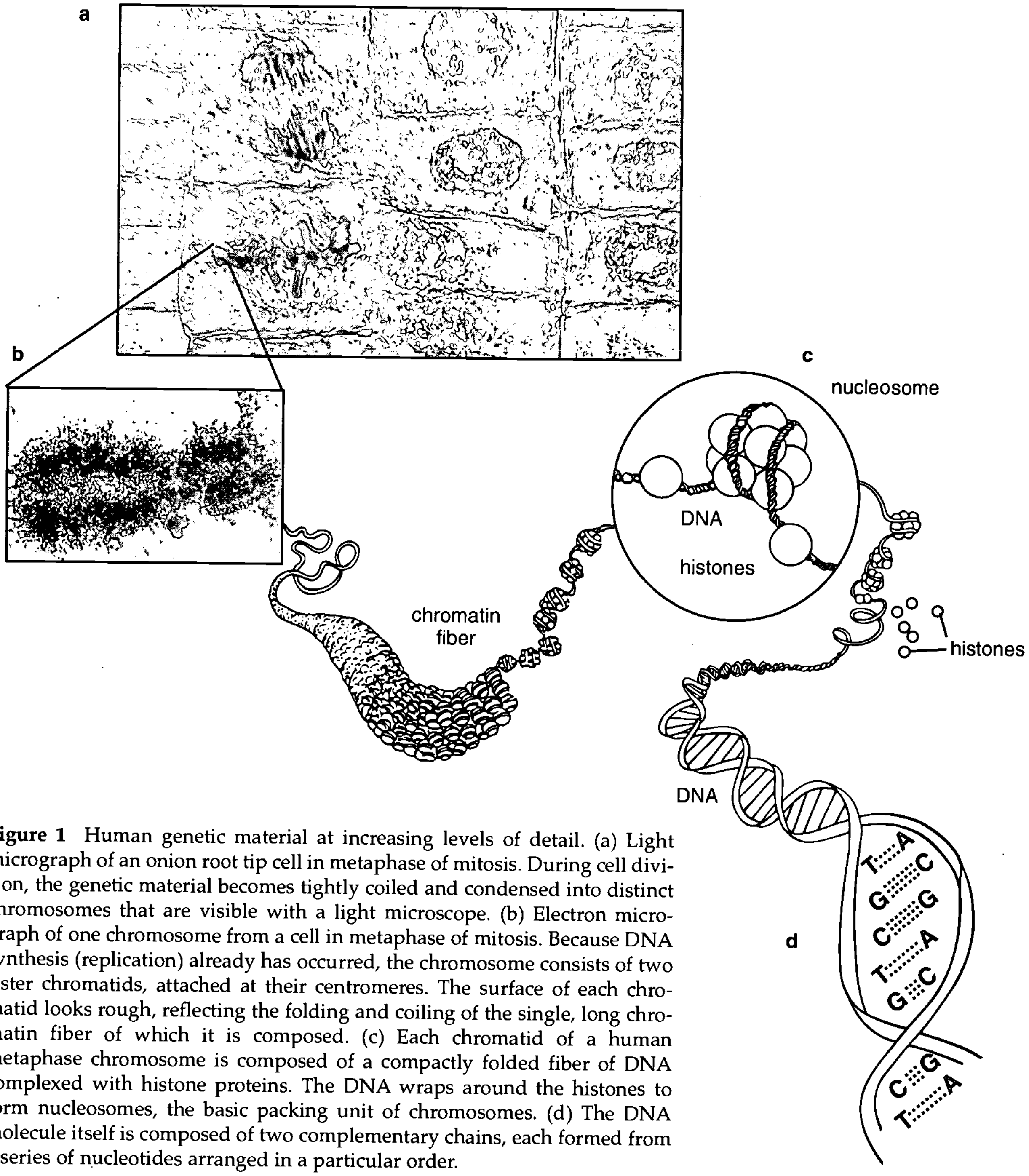


Figure 1 Human genetic material at increasing levels of detail. (a) Light micrograph of an onion root tip cell in metaphase of mitosis. During cell division, the genetic material becomes tightly coiled and condensed into distinct chromosomes that are visible with a light microscope. (b) Electron micrograph of one chromosome from a cell in metaphase of mitosis. Because DNA synthesis (replication) already has occurred, the chromosome consists of two sister chromatids, attached at their centromeres. The surface of each chromatid looks rough, reflecting the folding and coiling of the single, long chromatin fiber of which it is composed. (c) Each chromatid of a human metaphase chromosome is composed of a compactly folded fiber of DNA complexed with histone proteins. The DNA wraps around the histones to form nucleosomes, the basic packing unit of chromosomes. (d) The DNA molecule itself is composed of two complementary chains, each formed from a series of nucleotides arranged in a particular order.

Techniques Used to Map Chromosomes

As discussed in Section I, the primary objective of the HGP is to create detailed genetic and physical maps of each of the human chromosomes. Without these maps to tell molecular biologists where to look for specific genes, the full nucleotide sequence will be largely uninterpretable. Generating a set of detailed maps for even a single chromosome, however, is not a trivial task. One way to understand the complexity involved is to consider some of the techniques that scientists use to construct these maps.

Constructing a genetic map by classical linkage analysis. Geneticists long have used a technique called *linkage analysis* to determine how frequently different forms of two variable traits are inherited together (that is, are not separated by recombination during meiosis). In general, the closer the two genes are on a given chromosome, the more likely it is that they will be inherited together.

One of the methods that geneticists use to identify sets of genes that belong to the same linkage group (sets of genes that are located close enough together on the same chromosome that they tend to be inherited together) is *pedigree analysis.* Studying family pedigrees allows investigators to trace genes as they move from one generation to the next, and, in some cases, to demonstrate a close association between a gene of interest and another "marker" gene that has an identifiable phenotype. To be useful for tracing inheritance, the genes involved must vary between individuals. For example, a parent might have form G of a given gene on one chromosome of a pair and form g on the other, making it possible to determine which of the two forms was passed on to his or her child. Linkage analysis also requires pedigrees of informative families, that is, families that have multiple affected individuals and in which certain family members are heterozygous for markers that are linked to the gene in question. Without these heterozygous family members, it is impossible to distinguish the chromosome that carries the allele of interest from its homologous counterpart. Figure 2 illustrates the inheritance of a dominant allele in relation to a tightly linked marker allele.

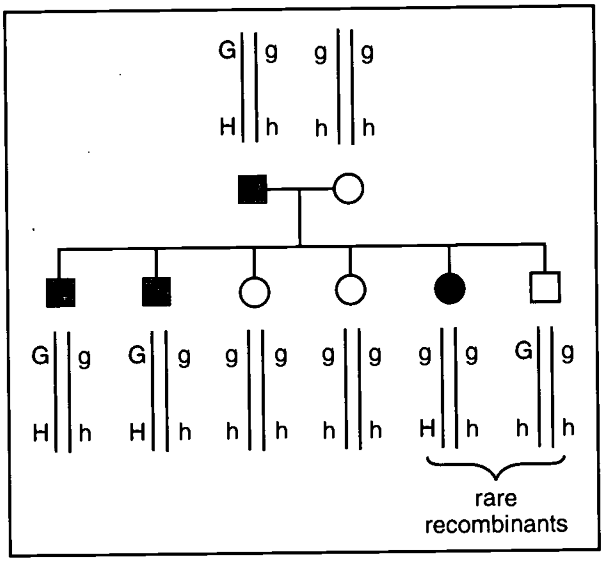


Figure 2 Inheritance of a dominant allele in relation to a tightly linked marker allele. G is an abnormal dominant allele associated with a genetic disease that is expressed late in life; g is its normal recessive counterpart. H is a dominant allele associated with a phenotype that always is expressed; h is its recessive counterpart. The frequency of coinheritance of G and H is related to the relative distance between them. If G and H lie very close together (that is, if the two genes are very tightly linked), then there is a high probability that G and H will be inherited together, and the presence of H indicates that G probably also is present. Therefore, the "H" gene acts as a marker for alleles of the "G" gene. (Individuals indicated by solid symbols display the trait associated with the H allele.)

Linkage analysis can help geneticists determine a gene's approximate chromosomal location without cloning the gene. If a gene is coinherited with a specific chromosomal feature (for example, an altered staining pattern), then one can assume that the gene is located near the area of the chromosome at which the feature appears. This information, in turn, helps scientists assign approximate locations to other genes known to be linked to the first. In this way, scientists can assemble *genetic maps* that indicate the order and relative distances between the genes that are located on specific chromosomes.

Constructing a genetic map by linkage analysis of polymorphic DNA markers. Although classical linkage analysis allows us to describe the relative distances between genes on a chromosome, it does not, in itself, provide the information required to identify exactly which piece of DNA carries a particular gene. This is because while a linkage study provides a *relative* measure of the distance between two genes, for various reasons, this relative measure is not always proportional to the *actual* physical length of the DNA involved.

Conducting the same type of linkage analysis by tracing *inherited variations in the base sequence itself,* however, can provide the missing information. A *DNA polymorphism* is a region within the DNA of a chromosome in which the base sequence varies from one person to another. Such a variable region along with the DNA probe that allows one to distinguish the various sequences is called a *polymorphic DNA marker*. Polymorphic DNA markers are powerful tools in mapping efforts for a number of reasons.

□ First, because polymorphic DNA markers vary among individuals and because scientists can detect these variations, we can trace their inheritance patterns much as we would trace the inheritance of a variable phenotypic trait. Scientists use this approach to create linkage maps (genetic maps) that illustrate the relative distance from one polymorphism to the next on the same chromosome.

□ Moreover, it also is possible to trace the coinheritance of a particular DNA marker and a variable phenotypic trait. This allows geneticists to calculate the relative distance (the genetic distance) from the marker to the gene for that trait.

□ Finally, because polymorphic DNA markers correspond to specific base sequences that can be recognized *at a molecular level,* scientists can determine each marker's actual physical location on a chromosome. This means that polymorphic markers can serve as landmarks in a search for the piece of DNA that contains a gene. If we know from the genetic map that a gene for a particular trait lies between two polymorphic DNA markers, and if we can find each marker's actual physical position on the DNA, then we also know that the gene lies on the piece of DNA contained between them. In other words, such DNA markers provide a way to connect the relative positions of genes and markers as they are shown on a genetic map with their actual physical positions on the DNA.

Figure 3 illustrates how linkage analysis using a polymorphic DNA marker can help investigators determine the physical location of a gene of interest. The importance of such markers to the HGP cannot be overestimated. A complete genetic map of human DNA with closely spaced markers will be crucial to the effort to locate and sequence all human genes. As described on p. 13, it was the discovery by linkage analysis that two different DNA markers appeared to flank the gene for cystic fibrosis that eventually led to this gene's isolation.

Constructing a physical map. Another key goal for the HGP is to create a detailed physical map of each human chromosome. A physical map illustrates the actual physical relationship between two points on a chromosome without regard to the frequency of their coinheritance.

One type of physical map that is particularly important to the HGP consists of an ordered set of contiguous overlapping DNA fragments that

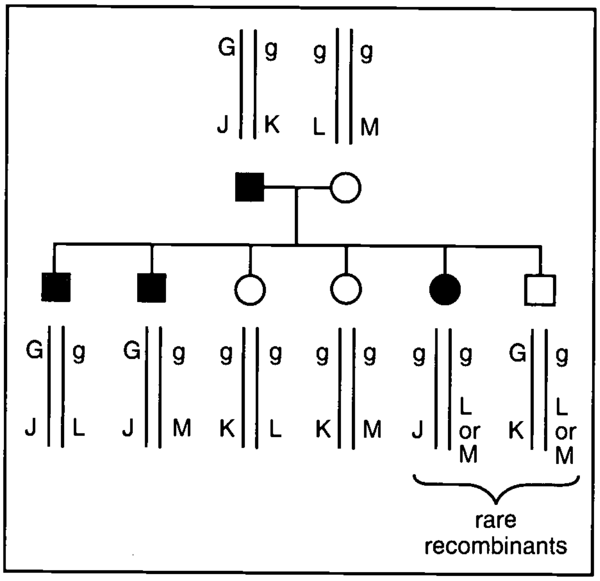
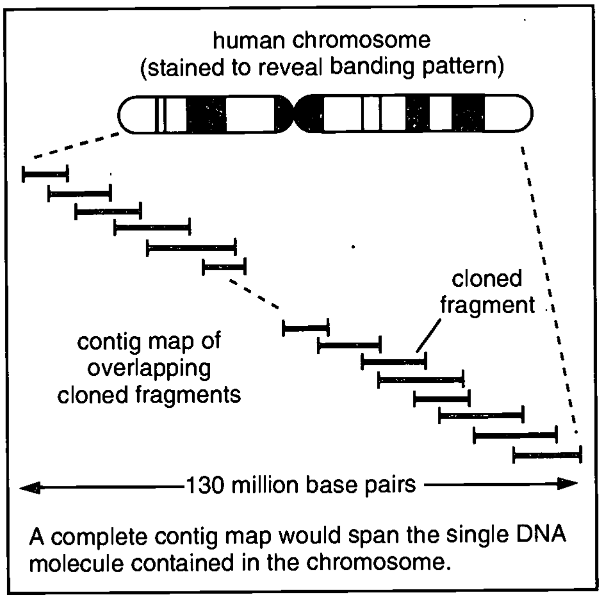


Figure 3 Inheritance of a dominant allele in relation to a tightly linked DNA polymorphism. This is the same pedigree that is illustrated in Figure 2, except that the marker in this case is a region on the chromosome that has a variable base sequence that is detectable using specific probes. Of the four possible base sequences in this region (J, K, L, and M), J is associated with the presence of allele G. That is, J serves as a marker for the presence of G. (Individuals indicated by solid symbols carry the J sequence.)

spans an entire section of a chromosome, or, even, the full length of the chromosome. Such a map, called a *contig map*, is a very difficult puzzle to assemble. To create such a map, investigators fragment the DNA under study using restriction endonucleases, enzymes that cut DNA molecules at sequence-specific points. This process, however, generates a set of fragments in no particular order, and scientists must use a variety of additional techniques to clone, analyze, and eventually to reorder the fragments (that is, to determine their original order along the chromosome). Ultimately, a contig map consists of a set of cloned, identified pieces of DNA that investigators can study in more detail. Figure 4 illustrates a contig map.

One importance of contig mapping to the HGP is that it produces libraries of cloned DNA fragments that have been identified as arising from specific sections of a chromosome. If scientists can map a polymorphic DNA marker to one of these ordered pieces (for example, by using the probe associated with that marker to screen the library), then they know the marker's approximate physical location along the chromosome. Equally important, they also know which piece of DNA in the library carries the marker. As noted above, if they have learned from linkage analysis that their marker and a gene of interest are located very close to each other, this piece of DNA also may carry the gene and may be a good candidate for more systematic and detailed study. Currently, the HGP includes research efforts to develop physical maps of all of the human chromosomes.


human chromosome
(stained to reveal banding pattern)
cloned fragment
contig map of overlapping cloned fragments
130 million base pairs

A complete contig map would span the single DNA molecule contained in the chromosome.

Figure 4 Example of a contig map. Assembling a complete contig map of even a small section of a chromosome can require investigators to analyze thousands of cloned DNA fragments. Scientists estimate that generating the yeast contig map required about 20 person-years of work. In contrast, even with recent improvements in technology, the complete physical mapping of each human chromosome will take about 100 person-years.

Techniques Used to Isolate and Identify Genes

As described in Section I, one important result of the research associated with the HGP will be the isolation and detailed study of many known human genes as well as the discovery and isolation of many as-yet-unknown genes. This raises two questions that are at the heart of the genome initiative: how do you find a gene and how do know it when you see it? If the human genome contains three billion nucleotides, how do you even begin to look for a gene?

The specific answers to these questions vary with the gene involved. The searches for some genes such as the human globin genes were helped by the ready availability of their mRNAs. Scientists can use these mRNAs as templates for the construction of cDNA (*complementary DNA*), which then can be used as a probe with which to locate the gene in a library of cloned DNA fragments. Other searches like the search for the gene for Duchenne muscular dystrophy begin with the observation that the condition invariably is associated with a particular chromosomal abnormality. In the case of DMD, investigators found that all affected females had translocations involving chromosome band Xp21. This relationship intensified efforts to examine DNA from this portion of the X chromosome to find the gene for DMD.

Despite differences in details, there are several major techniques that researchers use to search for genes. These include the mapping procedures described above, as well as a variety of specialized techniques that allow researchers to scan the DNA that lies in both directions from an identified marker for the presence of the gene.

One procedure that researchers use to characterize a region of the DNA and to locate a gene is called *chromosome walking* (Figure 5). Chromosome walking begins with a probe that recognizes a marker known to be closely linked to the gene of interest. Scientists use this probe to screen a library of cloned DNA fragments to find a fragment that overlaps a portion of the probe. They then use this fragment as a second probe to screen the same library, and so on, with each probe identifying the next piece of DNA along the chromosome. Thus, a successful chromosome walk identifies a series of partially overlapping fragments that, theoretically at least, contain DNA that is closer and closer to the gene of interest. Scientists keep track of where they are along the chromosome by checking the fragments for other known markers and by scanning the fragments' sequences for indications of the presence of genes.

Central to the usefulness of chromosome walking is the investigator's ability to recognize a gene when he or she reaches it. One way to do this is to look for sequences that show similarities to other sequences of known function, even those that may occur in other organisms. These similar sequences, which have been conserved through common evolutionary histories, help investigators distinguish DNA that is likely to represent a functional area from surrounding noncoding DNA.

Another way to find a gene in newly sequenced DNA is to search for stretches of DNA that are not interrupted by stop codons. Recall that in general, a gene contains the information that a cell needs to make a particular protein. For a cell to use this information, the DNA sequence first must be transcribed into RNA, which then is translated on ribosomes into protein. The genetic code is made

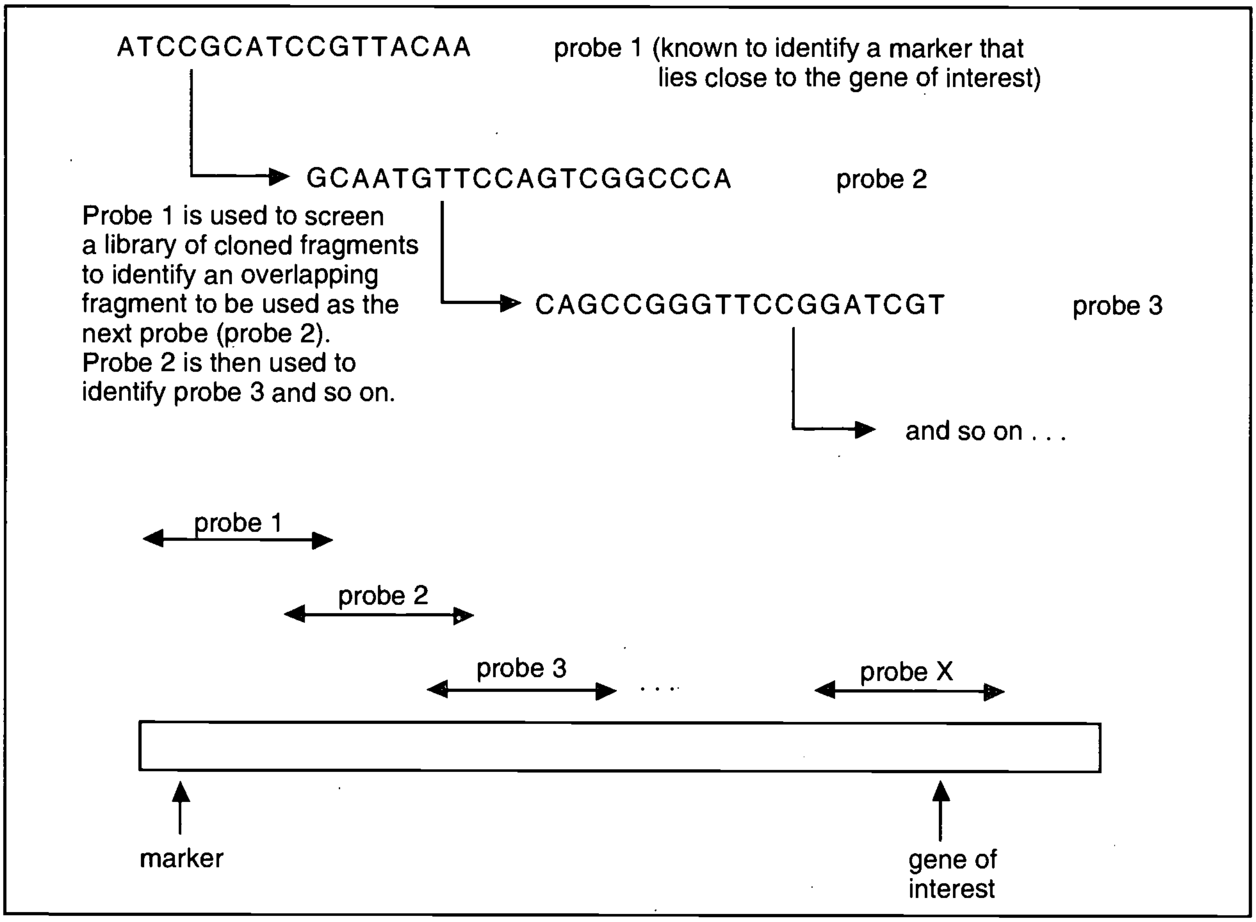

ATCCGCATCCGTTACAA
probe 1 (known to identify a marker that lies close to the gene of interest)
GCAATGTTCCAGTCGGCCCA
probe 2
Probe 1 is used to screen a library of cloned fragments to identify an overlapping fragment to be used as the next probe (probe 2). Probe 2 is then used to identify probe 3 and so on.
CAGCCGGGTTCCGGATCGT
probe 3
and so on . . .
probe 1
probe 2
probe 3
. . . .
probe X
marker
gene of interest

Figure 5 Chromosome walking. Unfortunately, executing a long chromosome walk can be a slow and tedious process, and sometimes the walk can be blocked by various technical problems. Nevertheless, as illustrated by the story of the search for the cystic fibrosis gene (p. 13), chromosome walking remains one of a gene hunter's most important tools.

up of "words," or *codons* that are three nucleotides long, most of which specify one of the 20 different amino acids found in protein. Three of the 64 possible codons, referred to as stop codons, tell the ribosome to stop making protein. The sequence of DNA or RNA between stop codons is called an *open reading frame (ORF)* because it would be possible for the ribosome to read all of the codons in that region if it found a start translation signal in that frame.

Because stop codons occur at random about every 20 bases along the DNA, the occurrence of a long ORF should be a good indication that a gene is present. Unfortunately, the correlation between long ORFs and genes is true only for bacteria and some other life forms. In higher organisms, the parts of a gene that code for protein (the exons) are interrupted by DNA sequences that do not code for protein (the introns) that the cell cuts out of the messenger RNA molecule before the RNA is translated. This situation makes it difficult to find genes in humans by inspecting the DNA sequence. Moreover, eukaryotic genomes often contain long open reading frames that are not expressed. Recently, however, scientists have found a set of properties of coding regions of human DNA that differ from those of noncoding regions. For example, certain sequence patterns that appear in coding regions do not seem to occur in noncoding regions. Scanning for these patterns can help investigators find genes in human DNA of unknown function.

Once investigators find a fragment of DNA that they believe contains the gene they are seeking, they must identify its coding regions and confirm its identity. An important step in this process usually is to probe the fragment using cDNAs prepared from the mRNA of cells in which the gene is expressed. If the identified fragment actually contains the gene, then one or more of the cDNAs should bind to it. This step also can help researchers locate the gene's exons. Once these coding regions are found, researchers can begin to sequence the gene in earnest, looking for differences between DNA derived from individuals known to carry the associated trait and DNA from individuals who do not carry it.

The Race to the Gene for Cystic Fibrosis

The search for the cystic fibrosis (CF) gene ended in 1989 with its discovery by Lap-Chee Tsui of Toronto's Hospital for Sick Children and Francis Collins at the Howard Hughes Medical Institute at the University of Michigan. At times highly competitive, at other times, highly collaborative, the search illustrates the combination of technical skill, ingenuity, and high-energy determination that sometimes is required to find a gene.

Locating the CF gene was a difficult challenge. First, despite years of study, investigators had little information available about the protein product that it encoded. Second, the gene was not associated with any apparent chromosomal change. In the absence of such clues, searching for the gene meant searching through the entire genome for a target that one could not describe in any detail.

Researchers took an important step forward when they finally mapped the gene to the long arm of chromosome 7, band q3. This victory was accomplished in a series of small steps as linkage analysis revealed first, a link between CF and a polymorphic DNA marker that mapped to chromosome 7; second, a link between CF and two other markers, both located on the long arm of chromosome 7; and third, after seven independent teams pooled their linkage data on more than 200 families, the critical observation that the two long-arm markers appeared to flank the CF gene (Figure 6).

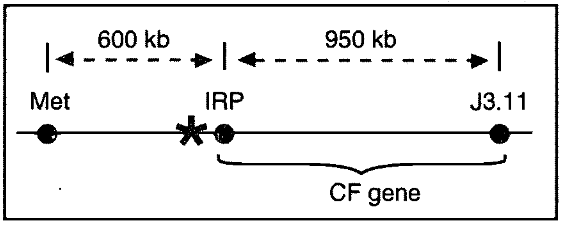


Figure 6 The search for the CF gene. Pooled data from seven independent research teams revealed that the cystic fibrosis gene was located somewhere between the IRP gene and J3.11, but closer to IRP. The asterisk marks the location at which the successful hunt for the gene began.

The discovery of flanking markers was an exciting breakthrough because now workers could focus their search on one region of the genome. Nevertheless, the piece of DNA between the two markers still was very long: as one research report phrased it, "a whopping 1.6 million bases long, long enough to hold as many as 50 to 100 genes."

Delighted as he was with the progress that the many labs searching for the gene had made, Lap-Chee Tsui still estimated that a chromosome walk along the indicated fragment would take an average lab about 18 years. Tsui's personal response to the challenge was, in his own words, to take a "brute force" approach that involved bombarding chromosome 7 with a large number of additional DNA probes in an effort to find at least one more marker that mapped closer to the CF gene than *either* of the two flanking markers. He screened more than 250 markers before he found two that looked promising.

At this point, Tsui joined his efforts with those of Francis Collins, who had developed *chromosome jumping*, a technique that allows investigators to skip over lengthy segments of DNA. Chromosome jumping speeds up a chromosome walk and avoids potential problems with unclonable or uninterpretable sequences. Even using this technique, however, the march to the gene crossed some 280,000 bases and took more than a year and a half. Initially, of course, investigators had to start at one of their new markers and walk and jump in both directions until they crossed a landmark that told them whether they were going toward the CF gene or away from it. When they found the IRP gene, a gene that another lab had identified as the closest marker yet to the CF gene, they knew they were going in the right direction (Figure 7).

As the walk proceeded, the scientists searched for potential genes by completing *zoo-blots* on each fragment. This technique compares the base sequence of a DNA fragment to the base sequences of known genes from a variety of animal species. The presence of structurally related sequences in other species suggests that the region has an essential function. The teams found three such conserved sequences, the first two of which they quickly ruled out as candidates for the CF gene. The third sequence, found to match sequences from a gene in chickens, mice, and cows, proved to be the beginning of the gene. Scientists were able to demonstrate that this gene is expressed in cells known to be affected by CF and that three of the gene's 250,000 bases are missing in most individuals with the condition. This small deletion leads to the loss of just one amino acid out of the 1,480 in the gene's protein product, but this loss is sufficient to disrupt radically the function of the individual's lungs, sweat glands, and pancreas.

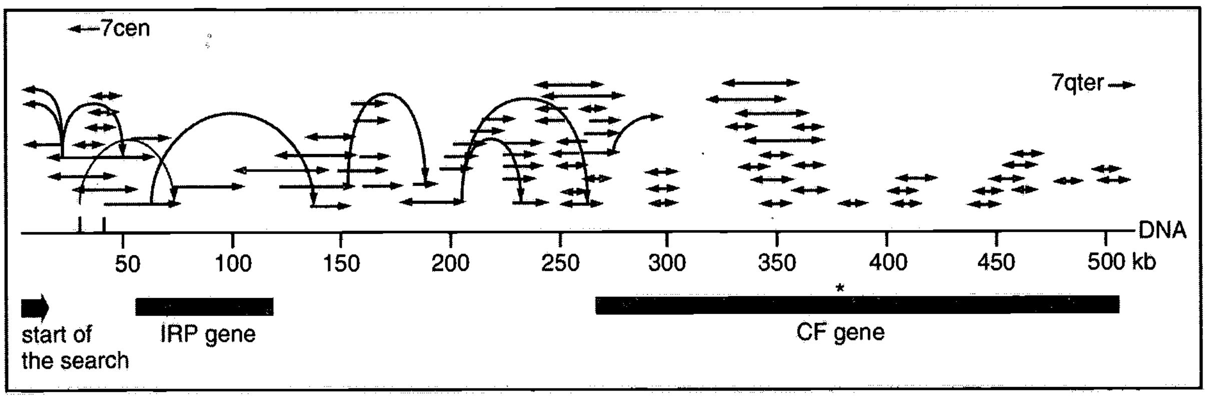


Figure 7 Walking and jumping to the cystic fibrosis gene. The trek began at a site, shown here at the left of the diagram, that Lap-Chee Tsui's group had identified as close to the CF gene on human chromosome 7. The 280 kilobases of DNA between the start of their search and the beginning of the gene were covered by a combination of chromosome "walking" and "jumping." The straight arrows above the line represent the DNA segments cloned during the walk and the curved arrows represent the jumps. The long bar on the lower right depicts the CF gene, which spans approximately 250 kilobases. The asterisk marks the location of the mutation that is found in 70 percent of abnormal CF genes.

The Role of Electronic Databases in the Human Genome Project

Research scientists predict that the effort to map and sequence all human genes will generate more data than any other single research effort in the history of biology. Simply to record the complete sequence of the haploid human genome will require the equivalent of 200 telephone books of 1,000 pages each. Variation in the sequences among humans as well as information about the genomes of other organisms will increase even further the amount of data that must be stored.

The most basic information that the HGP generates is the location of genes and other markers on chromosomes. Additional information about genes that is being generated includes the following:

□ variations in the structures and sequences of genes and markers among individuals (*polymorphisms*);

□ methods for detecting genes and markers and their variations using special probes and cell lines;

□ instructions for preparing DNA probes, the functions and characteristics of probes, and where to obtain them;

□ amino acid sequences of the protein products of genes, functions of proteins, and the possible three-dimensional structures of proteins;

□ clinical information about genetic diseases;

□ patterns of inheritance of genes through families (pedigrees); and

□ bibliographic references for all of the data.

As the volume and the complexity of this information have increased, the technologies required to store, manage, and communicate it also have changed. We can illustrate this relationship by comparing the increase in data during the last 20 years with changes in the technologies that scientists have used to store and manipulate these data (Figure 8).

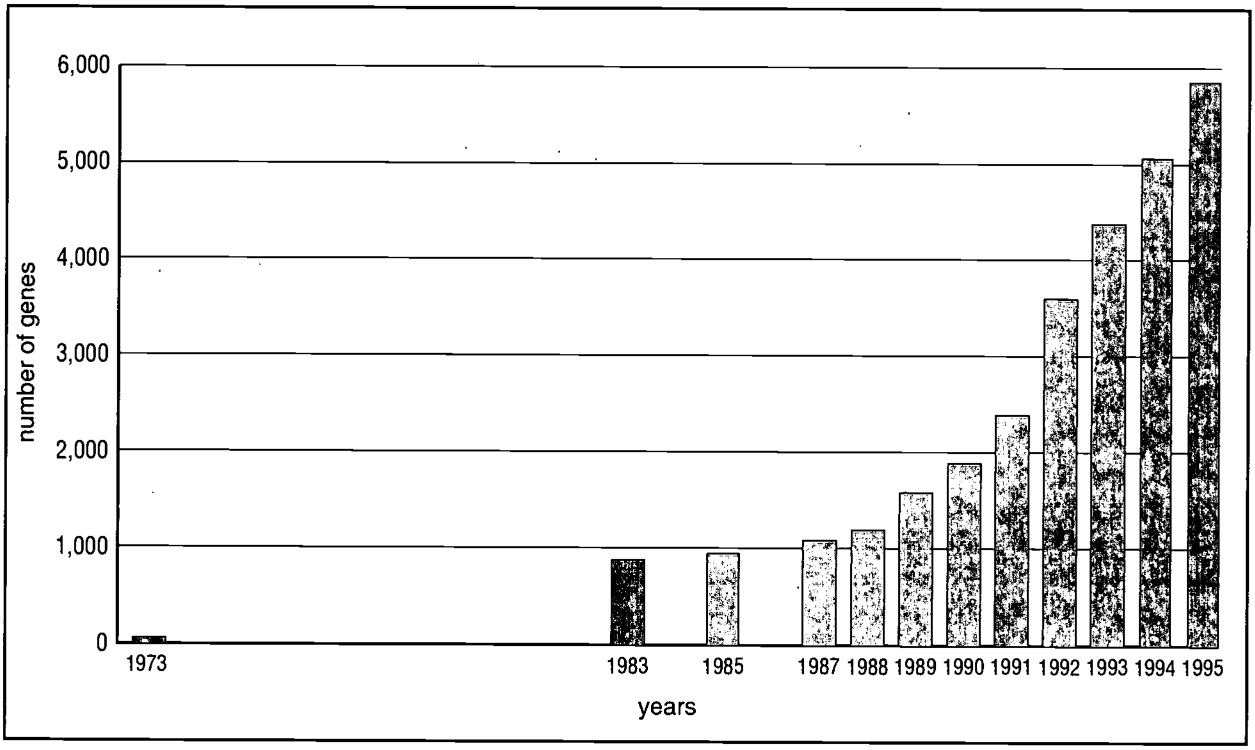


Figure 8 The increase in human genes involved in mapping studies from 1973 to 1995. The number has increased rapidly in the last 20 years. Likewise, the technologies used to manage these data have changed. In 1973, information about the 25 human genes involved in mapping studies could be summarized on one printed page; in 1995, the accumulated information on the more than 6,000 such genes was stored and made accessible through centralized, electronic databases. Data from GDB (Genome Database).

Electronic databases. A *database* is any collection of information that is organized so that people easily can find a particular piece of information. For example, a box of index cards with recipes from all over the world arranged by the name of the dish would constitute a very simple database. If you also grouped the recipes by the type of dish (for example, meat, fish, or desserts) and arranged each group by name, your database would be a little more complex.

Although a print-based database can be a useful tool, especially when the volume of information is small, such systems have several important limitations. For example, suppose you wanted to use your database to find all of the dishes from a specific country. This would not be easy if your index cards—your "database"—were arranged by the type of dish. To make this new type of search easy and efficient, you would have to create a duplicate set of cards and arrange *these* cards by the country. Clearly, the more information you want to store, and the more ways you want to be able to search through it, the more complicated the problem of organizing your cards becomes. A card catalog in a library is another example of a print-based database; note that a card catalog must be duplicated three times to allow a user to search by author, title, *and* subject.

Many of the disadvantages of using print-based databases to store and organize large volumes of information are eliminated by using *computers* to accomplish these tasks. Computer programs that can collect, organize, store, and display data are called *database software*. Together, a computer, the data, and the software required to store, manipulate, and display those data make up an *electronic database*.

Without electronic databases, the research community would be unable to collect, store, organize, or use HGP-related data effectively. Electronic databases require less space than most print-based systems, can be kept consistent and accurate more easily, and allow searches that automatically identify all relevant records as determined by multiple criteria. *Online databases* (databases that can be accessed using computers linked by telecommunications) offer the additional advantages of immediate (real time) updating, multiple users, and access to data that are stored in many remote locations.

These advantages are vital to the HGP. The first nucleotide sequence, that of an RNA molecule fewer than 100 nucleotides long, was published in 1965. By 1996, investigators had published more than 460 million nucleotides of sequence data, as well as millions of bits of other types of genome-related information. The enormous volume of the information that must be stored and the need to search these data rapidly and efficiently preclude using a print-based medium, which would have to be maintained separately in each research location, would occupy hundreds of pages, and could be searched only in very prescribed and limited ways. In contrast, real time updates of an online system, often completed by the actual lab that generated the data, make new research findings available to other scientists quickly and accurately. This process avoids publication delays and also avoids errors that can be introduced as data are transcribed to create published accounts. The ability to access data that are stored in remote locations allows scientists around the world to retrieve and use a wealth of different types of information.

Currently, genomic data are stored in two types of databases, *research databases* and *registries*. As described below, these two types of databases are distinguished by their content, purposes, and accessibility.

Research Databases and the Human Genome Project

Research databases store the results of scientific research and are used to support continued research and to improve the quality of patient care. Many research databases are *private* databases, maintained entirely within one laboratory or one group of collaborating laboratories. Scientists use these databases to collect, store, and analyze the immediate results of research (the raw data). In contrast, other research databases are *public* databases; public databases act as central collection sites for experimental results that are sufficiently sound and complete that scientists are

Table 2 The volume of information stored in three major research databases.

Database	Type of Information	Number as of January 1996	Increase Since January 1995
GDB (Genome Database)	genes	5,902	733
	markers	78,184	22,896
	polymorphisms	17,226	3,833
	probes	313,026	156,541
	references	56,103	7,164
GenBank	number of nucleotide sequences	685,693	416,215
	number of bases in all sequences	463,758,833	215,259,619
OMIM (Online *Mendelian Inheritance in Man*)	number of genetic traits, including disorders	7,980	800
	number of traits mapped to autosomes (1-22)	7,435	751
	number of traits mapped to sex chromosomes (X,Y)	485	49

willing to share them with other research facilities around the world. These public databases are available to anyone who wants to access them for reading, although there are various limitations on who can access them to enter or change information. Table 2 illustrates the volume of genomic information that was stored in three major research databases at the beginning of 1996, as well as the increase in these data that occurred between January 1995 and January 1996. Table 3 summarizes the content and purposes of the major public databases, and the sections below describe some of the ways in which scientists use these databases.

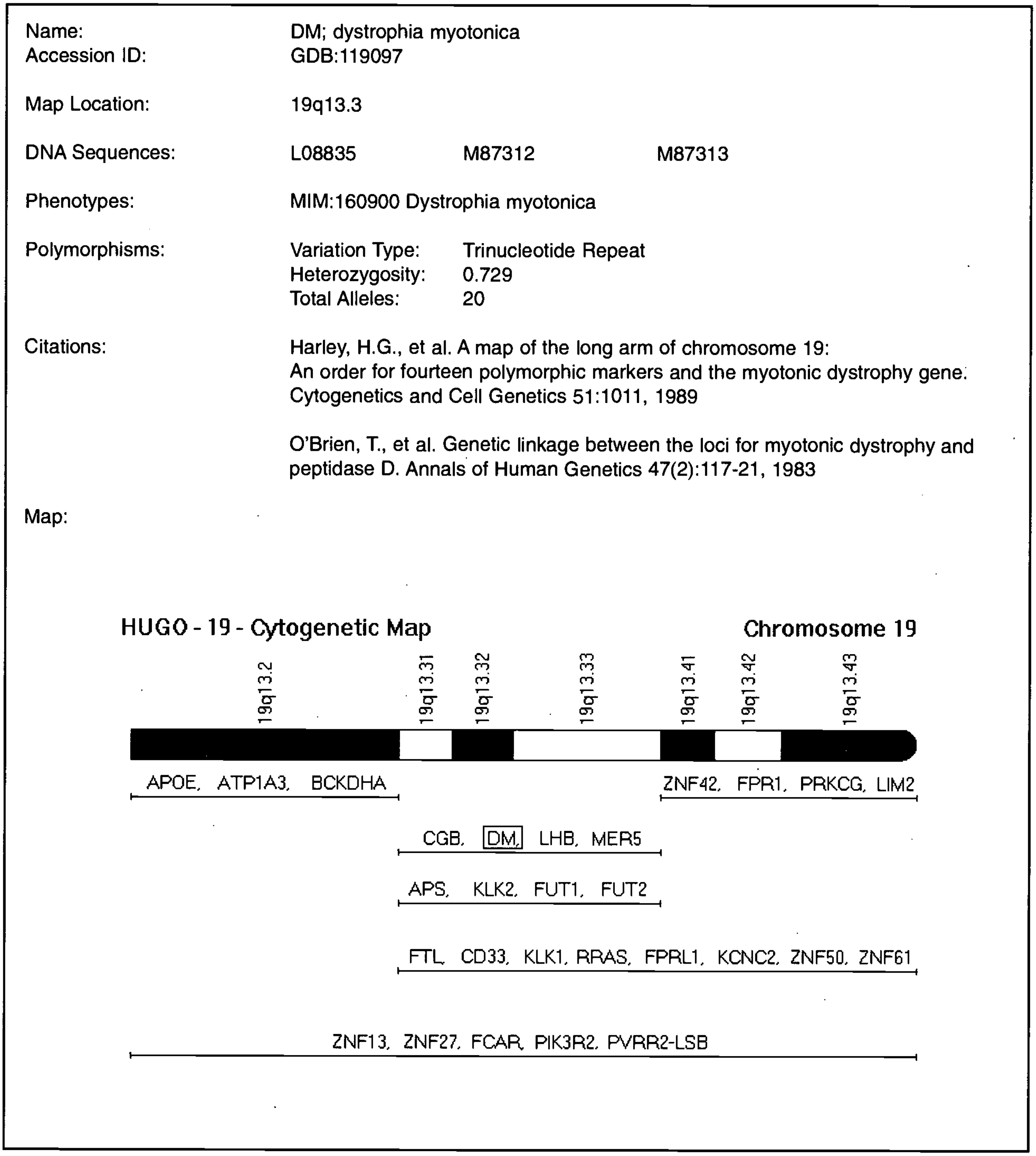
Map data. Mapping databases contain information about the location of genes and markers on chromosomes. The Genome Database (GDB), located at Johns Hopkins University in Baltimore, Maryland, stores human mapping data. The data in GDB specify the position of a given gene or marker as precisely as possible, including the location of the segment of DNA on the chromosome, the relative order of surrounding genes and markers, a genetic measure of the distances between them, and bibliographic references for all the data. For example, Figure 9 shows an edited record of the myotonic dystrophy gene as it is represented in GDB. This record shows that the gene for myotonic dystrophy is located on the long arm of chromosome 19 (19q), in region 1, band 3, sub-band 3. The record also gives the names for the gene (DM, dystrophia myotonica), indicates that the gene is polymorphic, and identifies the type of variation that has been found among different alleles (trinucleotide repeat).

The ability to retrieve the chromosomal location of a gene from a mapping database allows scientists to ask and answer a variety of questions. For example, sometimes investigators know from pedigree analysis the approximate location of a disease gene. A scientist interested in this gene can consult a mapping database to see whether any gene of known biochemical function has been mapped to the same neighborhood. The investigator then can ask, using information in the database about the phenotype of the disease

Table 3 An overview of selected research databases.

Database	Type of Data	Location	Sample Question that Could Be Answered Using the Database
GDB (Genome Database)	map data	Johns Hopkins University, Division of Biomedical Information Sciences (Baltimore, Maryland)	What genes are located on the short arm of chromosome 7?
GenBank	sequence data	National Center for Biotechnology Information at the National Library of Medicine (Bethesda, Maryland)	Is anything known about this piece of DNA that we just sequenced?
GSDB (Genome Sequence Data Base)	sequence data	National Center for Genome Research (Santa Fe, New Mexico)	Is anything known about this piece of DNA that we just sequenced?
OMIM (Online *Mendelian Inheritance in Man*)	clinical descriptions	Center for Medical Genetics, Johns Hopkins Hospital (Baltimore, Maryland)	What clinical symptoms are associated with fragile X syndrome?
PIR (Protein Information Resource)	protein structure data	National Biomedical Research Foundation (Washington, D.C.)	Have any proteins with a structure similar to this been identified and characterized?
POSSUM (Pictures of Standard Syndromes and Undiagnosed Malformations)	diagnostic data	CD-ROM (product used at investigator's location)	Do these clinical symptoms match any known genetic disorder?
ATCC (American Type Culture Collection)	supply data	Rockville, Maryland	Where can I obtain the cloned probes that were used to generate these map data?
MEDLINE	bibliographic data	National Library of Medicine (Bethesda, Maryland)	What are the latest published findings on Huntington disease?

gene, whether the known gene might be a candidate for the disease gene of interest. The gene associated with retinitis pigmentosa was discovered when an investigator determined through linkage studies that the condition seemed to be linked to a particular region of chromosome 3. A search of GDB revealed that a gene known to be important to vision is located on the same chromosome. When scientists analyzed the DNA in the area of this gene more closely, they discovered that its sequence in people with retinitis pigmentosa is different from its sequence in people without the condition. In this case, the scientists' ability to use a research database to connect mapping information generated in one laboratory to that generated in another allowed identification of a gene that otherwise might have remained unknown for years.



Name: DM; dystrophia myotonica
Accession ID: GDB:119097

Map Location: 19q13.3

DNA Sequences: L08835 M87312 M87313

Phenotypes: MIM:160900 Dystrophia myotonica

Polymorphisms:
Variation Type: Trinucleotide Repeat
Heterozygosity: 0.729
Total Alleles: 20

Citations:
Harley, H.G., et al. A map of the long arm of chromosome 19: An order for fourteen polymorphic markers and the myotonic dystrophy gene. Cytogenetics and Cell Genetics 51:1011, 1989

O'Brien, T., et al. Genetic linkage between the loci for myotonic dystrophy and peptidase D. Annals of Human Genetics 47(2):117-21, 1983

Map:

Figure 9 The myotonic dystrophy gene as represented in GDB [this edited record lists only some of the data for the DM gene]. Note the use of unique numbers to identify a record within a database and to provide cross-references to related data in other databases. The unique number of the myotonic dystrophy gene in GDB is GDB:119097. The related GenBank data are identified by the GenBank numbers L08835, M87312, and M87313 (see Figure 10), and the related OMIM data by the MIM number 160900 (see Figure 11). The graphic map illustrates the long arm of chromosome 19 from 19q13.2 to the telomere, showing the map positions of the DM gene and some of the other genes nearby. The lengths of the lines indicate how precisely the genes have been mapped. For example, DM and the other three genes on the same line (CGB, LHB, MER5) are located somewhere within the area of the chromosome designated 19q13.3. By comparison, the genes FTL, CD33...ZNF61 are located somewhere between 19q13.3 and 19q13.4 and the genes ZNF13, ZNF27...PVRR2-LSB are located somewhere between 19q13.2 and 19q13.4.

Map data from organisms other than humans also sometimes can help scientists map new human genes. This is because groups of genes often remain together (that is, are linked) during long periods of evolutionary time. For example, several genes on the X chromosome appear to have remained clustered throughout mammalian evolution. This tendency for sets of genes to be located together in the genomes of closely related organisms gives researchers important clues about where to look for similar genes in humans. In fact, because of the speed at which new genes can be mapped in the mouse and because of the information that scientists have collected already about similarities and differences between the mouse and human genomes, it often is easier to determine a new gene's location in the mouse and then to predict its location in humans than to map the gene in humans first.

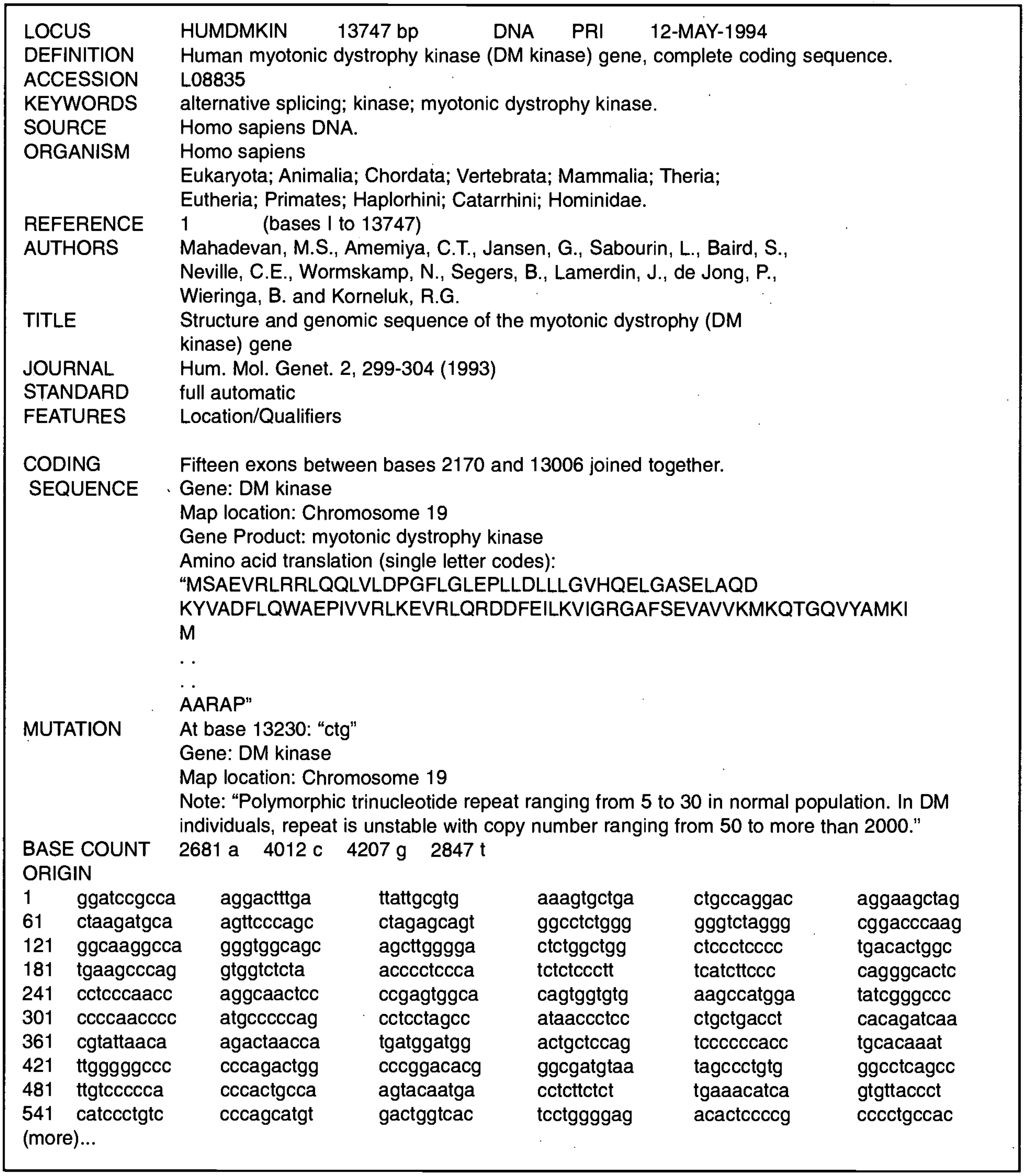
Sequence data. Sequence databases contain data about DNA, RNA, or protein sequences, including the sequences themselves and a variety of related information. Figure 10 shows an edited record for the myotonic dystrophy gene from GenBank, a DNA sequence database. This record shows the source from which the DNA was obtained (human DNA), the number of bases sequenced (13,747), and the full bibliographic reference where the sequence was reported. In addition, the GenBank record summarizes key features of the gene and lists its actual base sequence.

Usually a sequence database will store and make available all of the sequences that have been reported to it. This means that if two labs report slightly different sequences for the same gene, both sequences will appear in the database. In this way, sequence databases attempt to make all sequence data available for independent analysis and interpretation by the scientific community. In July 1993, GenBank contained three different records for the myotonic dystrophy gene, the record shown (L08835), which gives a full sequence, and two others. The two additional records illustrate the instability in the trinucleotide repeat area: one sequence shows five repeats and the other shows 11.

Scientists use sequence databases for a number of purposes. Some scientists use DNA sequence databases to locate genes. As discussed earlier, this may involve searching the database for sequences that show similarities to sequences of known function in other species, or it may involve analyzing the stored sequences for long ORFs or for particular sequence patterns that may signal the presence of a gene.

Likewise, scientists may use protein sequence databases to gain clues about the function of proteins predicted from the structures of newly sequenced DNA. An example of a protein sequence database is Protein Information Resource (PIR), located at the National Biomedical Research Foundation in Washington, D.C. One of the ways in which scientists use protein sequence information is to compare sequences of one species with those from another, or the sequence from one protein with that of another. Although structural similarities do not always result in functional similarities, insights gained through sequence comparisons often give researchers a place to begin in their effort to determine the function of an unknown protein. Information about the three-dimensional structures of proteins, also important for understanding protein function, is stored in the Protein Data Bank (PDB) at Brookhaven National Laboratory in Upton, New York.

Finally, detailed comparisons of the sequences of both nucleic acids and proteins can help scientists answer a wide range of questions about the evolution of life. A protein is a polymer of 20 different amino acids whose order is determined by the sequence of the nucleotides that make up the gene that encodes the particular protein. If proteins were random polymers of amino acids, the odds that two proteins would contain a sequence of five amino acids in exactly the same order would be less than one in three million. However, proteins are not random polymers. Their amino acid sequences not only determine their function but also reflect their evolutionary history. Proteins that share an identical sequence over hundreds of amino acids, as is the case between human and chimpanzee hemoglobins, show the degree of relatedness between the species. It is clear that this conservation of sequence between proteins of closely related species reflects more than just an evolutionary pressure to conserve the function of the protein (and therefore its sequence), because homologous proteins from more distantly related species do show differences in amino acid sequence, even though they all are functional proteins. The extension activity included in this module offers students an opportunity to explore the ways in which DNA and protein sequence data are important to the study of phylogeny and evolution.



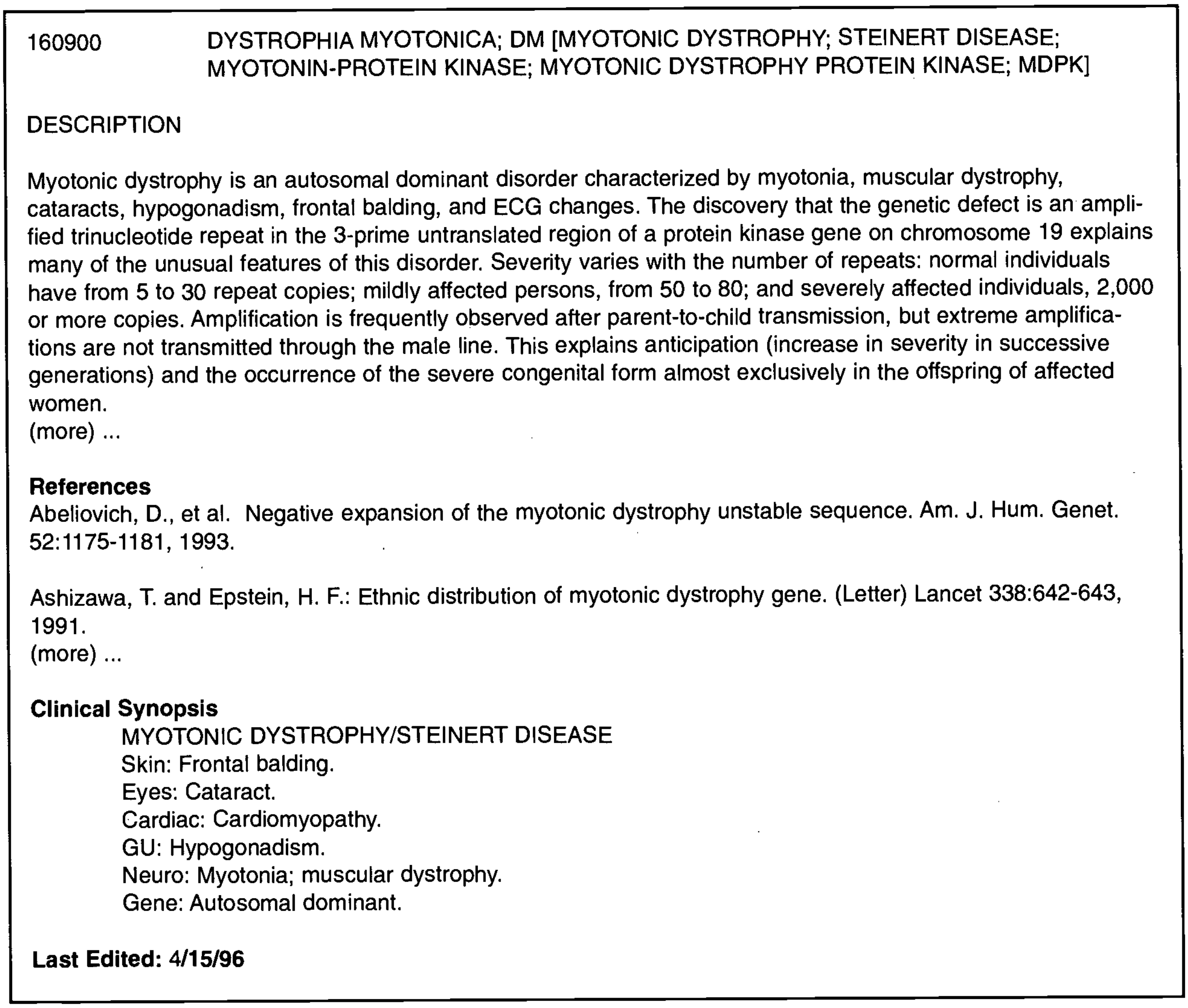
```
LOCUS          HUMDMKIN      13747 bp        DNA     PRI      12-MAY-1994
DEFINITION     Human myotonic dystrophy kinase (DM kinase) gene, complete coding sequence.
ACCESSION      L08835
KEYWORDS       alternative splicing; kinase; myotonic dystrophy kinase.
SOURCE         Homo sapiens DNA.
ORGANISM       Homo sapiens
               Eukaryota; Animalia; Chordata; Vertebrata; Mammalia; Theria;
               Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
REFERENCE      1        (bases l to 13747)
AUTHORS        Mahadevan, M.S., Amemiya, C.T., Jansen, G., Sabourin, L., Baird, S.,
               Neville, C.E., Wormskamp, N., Segers, B., Lamerdin, J., de Jong, P.,
               Wieringa, B. and Korneluk, R.G.
TITLE          Structure and genomic sequence of the myotonic dystrophy (DM
               kinase) gene
JOURNAL        Hum. Mol. Genet. 2, 299-304 (1993)
STANDARD       full automatic
FEATURES       Location/Qualifiers

CODING         Fifteen exons between bases 2170 and 13006 joined together.
 SEQUENCE      Gene: DM kinase
               Map location: Chromosome 19
               Gene Product: myotonic dystrophy kinase
               Amino acid translation (single letter codes):
               "MSAEVRLRRLQQLVLDPGFLGLEPLLDLLLGVHQELGASELAQD
               KYVADFLQWAEPIVVRLKEVRLQRDDFEILKVIGRGAFSEVAVVKMKQTGQVYAMKI
               M
               . .
               . .
               AARAP"
MUTATION       At base 13230: "ctg"
               Gene: DM kinase
               Map location: Chromosome 19
               Note: "Polymorphic trinucleotide repeat ranging from 5 to 30 in normal population. In DM
               individuals, repeat is unstable with copy number ranging from 50 to more than 2000."
BASE COUNT     2681 a    4012 c    4207 g    2847 t
ORIGIN
1      ggatccgcca    aggactttga    ttattgcgtg    aaagtgctga    ctgccaggac    aggaagctag
61     ctaagatgca    agttcccagc    ctagagcagt    ggcctctggg    gggtctaggg    cggacccaag
121    ggcaaggcca    gggtggcagc    agcttgggga    ctctggctgg    ctccctcccc    tgacactggc
181    tgaagcccag    gtggtctcta    acccctccca    tctctccctt    tcatcttccc    cagggcactc
241    cctcccaacc    aggcaactcc    ccgagtggca    cagtggtgtg    aagccatgga    tatcgggccc
301    ccccaacccc    atgcccccag    cctcctagcc    ataaccctcc    ctgctgacct    cacagatcaa
361    cgtattaaca    agactaacca    tgatggatgg    actgctccag    tccccccacc    tgcacaaat
421    ttgggggccc    cccagactgg    cccggacacg    ggcgatgtaa    tagccctgtg    ggcctcagcc
481    ttgtccccca    cccactgcca    agtacaatga    cctcttctct    tgaaacatca    gtgttaccct
541    catccctgtc    cccagcatgt    gactggtcac    tcctggggag    acactccccg    cccctgccac
(more)...
```

Figure 10 Edited version of the myotonic dystrophy gene as represented in GenBank.

Clinical data. Clinical databases contain descriptions of genetic disorders and traits, as well as information about how these traits are inherited, their chromosomal location (if known), and relevant publications. The best known clinical database is Online *Mendelian Inheritance in Man* (OMIM), located at the Center for Medical Genetics at Johns Hopkins Hospital in Baltimore, Maryland. OMIM was established in 1985 as a computer-based version of Victor McKusick's classic catalog of human genetics, *Mendelian Inheritance in Man*. The database contains general information synthesized from the literature about all known human genes and clinical disorders. Figure 11 shows an edited version of the OMIM entry for myotonic dystrophy. Access to OMIM is provided by the National Center for Biotechnology Information in Bethesda, Maryland.

Scientists use OMIM to look for descriptive summaries of specific conditions and to identify research that reports on the inheritance patterns of traits of interest. Genetic counselors, nurses, and physicians also use databases such as OMIM to locate clinical information about a variety of known genetic conditions. Because OMIM is available online, the data it contains often are more current than the data in any printed book or published journal could be. Clinicians using it and other similar databases usually can be sure that they have the latest information to share with their patients.



160900 DYSTROPHIA MYOTONICA; DM [MYOTONIC DYSTROPHY; STEINERT DISEASE; MYOTONIN-PROTEIN KINASE; MYOTONIC DYSTROPHY PROTEIN KINASE; MDPK]

DESCRIPTION

Myotonic dystrophy is an autosomal dominant disorder characterized by myotonia, muscular dystrophy, cataracts, hypogonadism, frontal balding, and ECG changes. The discovery that the genetic defect is an amplified trinucleotide repeat in the 3-prime untranslated region of a protein kinase gene on chromosome 19 explains many of the unusual features of this disorder. Severity varies with the number of repeats: normal individuals have from 5 to 30 repeat copies; mildly affected persons, from 50 to 80; and severely affected individuals, 2,000 or more copies. Amplification is frequently observed after parent-to-child transmission, but extreme amplifications are not transmitted through the male line. This explains anticipation (increase in severity in successive generations) and the occurrence of the severe congenital form almost exclusively in the offspring of affected women.
(more) ...

References
Abeliovich, D., et al. Negative expansion of the myotonic dystrophy unstable sequence. Am. J. Hum. Genet. 52:1175-1181, 1993.

Ashizawa, T. and Epstein, H. F.: Ethnic distribution of myotonic dystrophy gene. (Letter) Lancet 338:642-643, 1991.
(more) ...

Clinical Synopsis
MYOTONIC DYSTROPHY/STEINERT DISEASE
Skin: Frontal balding.
Eyes: Cataract.
Cardiac: Cardiomyopathy.
GU: Hypogonadism.
Neuro: Myotonia; muscular dystrophy.
Gene: Autosomal dominant.

Last Edited: 4/15/96

Figure 11 Edited version of the myotonic dystrophy gene as represented in OMIM.

Researchers and clinicians, likewise, often use diagnostic databases to locate specific information about the clinical symptoms that are associated with congenital defects. Congenital defects are abnormalities that are present at birth and include genetic disorders (for example, myotonic dystrophy or achondroplasia), chromosomal abnormalities (for example, Down syndrome), and developmental abnormalities (for example, fetal alcohol syndrome). Producers of these databases organize the entries to allow users to search records by particular symptoms (for example, short stature, cleft palate, or blindness). Searches of this type help clinicians make differential diagnoses of patients with unknown disorders. An example of a clinical database is Pictures of Standard Syndromes and Undiagnosed Malformations (POSSUM), a CD-ROM database that contains more than 25,000 illustrations of patients with various congenital disorders.

Access to Selected Research Databases via Internet the World Wide Web (WWW)[2]

Database	URL
Genome Database	http://gdbwww. gdb.org/
GenBank	http://www.ncbi.nlm.nih.gov/
Online *Mendelian Inheritance in Man*	http://www3.ncbi.nlm.nih.gov/Omim/

Other research databases. Together, GDB, GenBank, and OMIM contain much of the genomic data that are generated as a result of HGP-related research. Other genomic data also related to the HGP are of a very different type and are stored in yet other types of databases. For example, as the number of cloned probes and cell lines that are available as research tools has increased, it has become important that information about these tools also be broadly available to the scientific community. Supply databases contain information about biological supplies used in genetic experiments, including descriptions of what is available, the source of the materials, and information about ordering. The American Type Culture Collection (ATCC), located in Rockville, Maryland, is a major source of the cloned probes that scientists use in mapping experiments, and is an important supplier of a large variety of bacterial, fungal, algal, and viral strains. Likewise, the Coriell Cell Repository, located in Camden, New Jersey, provides extracted DNA and live cells from specific somatic cell lines used in mapping experiments. Similarly, bibliographic databases like MEDLINE, located at the National Library of Medicine in Bethesda, Maryland, contain references to the biological and biomedical literature worldwide and provide detailed descriptions of the research that has generated the data stored in the other types of databases.

As impressive as this list of databases is, it does not capture fully the magnitude of the data collection efforts currently in place. Nor does it reveal the difficulties that the scientific community already is facing with respect to the storage, manipulation, and communication of genomic data. One serious problem has to do with the flow of information from the research labs, where the data are generated, to the public databases. This flow depends heavily on mutual accountability and cooperation. For the most part, those responsible for maintaining databases such as GDB and GenBank must rely on research scientists to enter new data and to update existing data. Inevitably, this raises questions about the competing values of currency and accuracy: although most scientists want new data to appear in the public databases as soon as possible, the sooner a research group releases new findings, the less chance it has to check the data. Furthermore, many scientists struggle with commercial and patent concerns. When genomic data have potential commercial value, some research labs are reluctant to share them. Research groups funded as part of the HGP, however, are required to make new data publicly available within six months after their discovery.

Another problem that the scientific community faces relates to the analysis and interpretation of raw data, particularly those data generated by large laboratories or by networks of research groups working on particular research questions. As several investigators or even several different labs begin to collaborate on the same or related issues, they must create systems for sharing and communicating information. For example, at least seven different laboratories around the world collaborate on the study of Huntington disease. These labs must create mechanisms for rapid storage and analysis of raw data, as well as systems for integrating those data into coherent answers to their research questions.

As an illustration of the enormous volume of data that some labs have to manage, consider investigators at the Los Alamos Genome Center in Los Alamos, New Mexico, where the effort to map human chromosome 16 is centered. These scientists currently track the sizes and sources of more than 100,000 fragments of DNA and store information about more than seven million pairwise positional relationships among these fragments that are relevant to the emerging map. In a recent research report, a scientist working at the genome center summarized the data problem: "At Los Alamos we have accumulated more information on chromosome 16 than we ourselves can access in an easy fashion. It has turned out to be a bottleneck for us. The problem of sending 4,000 clones someplace is easy compared with sending the information we've accumulated on chromosome 16 in some useful and intelligible format."[3]

*Long Distance Research**

I support GDB and OMIM users located all over the world. We have had users register from such far-flung places as Russia and Saudi Arabia, but the most amazing interaction I have had so far is the Internet relationship that I developed with a doctor practicing in Malaysia.

One morning in June 1992, I checked my e-mail queue, which receives messages sent to our "help" address. One of the messages was from a Dr. Elizabeth Hillman, who described herself as "a Canadian pediatrician working in an isolated medical school on the underserved east coast of Malaysia." She went on to say that they had just successfully connected to Internet and had no trouble communicating.

I sent her back an e-mail telling her that we would be happy to register her. Word spread around the office and more than once the comment was heard "How isolated can they be if they are on the Internet?" We mailed her a letter welcoming her to the system and shipped her our documentation under separate cover. Before she received her documentation, however, she e-mailed us again: "I expect I should wait for my package but all my colleagues are so keen to get to using the database that I am being pressured into contacting you." Dr. Hillman then asked us to help her solve a medical mystery of a woman and her premature baby. She described the baby's condition in detail, and asked us to find out if it had a name, and whether the mother's future children could be affected.

Although we normally are too short-staffed to perform searches for users, we decided that she was an exception. My co-worker, Kerryn Brandt, searched OMIM, and forwarded three relevant entries. Dr. Hillman was able to diagnose the baby as having Baller-Gerold Syndrome, a relatively rare condition. Dr. Hillman responded with her thanks: "Do you know what something like this means in a really isolated place where such information is impossible to find?..." She mentioned wishing she could send us orchids by e-mail, but her grateful e-mail was the equivalent.

I heard from Dr. Hillman a few times over the summer, and I learned that she finally received our documentation. I also learned she was coming back to Canada for her daughter's wedding. I hope that one day we can meet, as she represents what electronic databases and user support are all about: people helping people.

We could not have done it without the Internet.

*by Patricia Haley, Documentation Coordinator/User Support Staff, Genome Database (GDB), Johns Hopkins University, Baltimore, Maryland.

Registry Databases and the Human Genome Project

Registry databases contain information identified as "belonging to" or being "descriptive of" specific people. That is, in contrast to research databases, which store generalized, anonymous information that is not connected to the specific individuals from whom the data were obtained, registry databases contain information that is specifically identified as describing or being derived from particular people. In fact, while the data in a research database generally have value and usefulness independent of specific information about the identities of the individuals from whom the samples were taken, the data in a registry have meaning and value largely *because* they are connected to the names of the people whom they describe.

In general, we can identify two types of registry databases. *Information databases*, such as those found in many government offices, businesses, and schools, are used as repositories of a wide variety of information about individuals. A computerized listing of students' grades in a college registrar's office constitutes a very straightforward registry database, as do the records accumulated by each state's motor vehicle and driver registration processes. By comparison, *identification databases* contain information that can be used specifically to help identify people. Examples of identification registries include the databases that are maintained by federal and state law enforcement agencies that contain records of people's physical characteristics such as weight, height, hair and eye color, skin marks, dental variations, and fingerprint patterns.

Information and identification databases usually store coded data rather than actual tissue samples. In contrast, databases that contain biological materials from which descriptive or identifying information could be derived—for example, samples of blood, saliva, body tissues, or even DNA—often are called databanks. We discuss databanks as a separate category of registry databases on pp. 29-31.

Because the information in a registry database is personal, access for entering or viewing the data usually is limited to individuals with appropriate authorization. Such limitations tend to be particularly stringent when the information is related to health. An individual's medical records, for example, are considered confidential and normally may not be accessed by a third party or released from a hospital or doctor's office without the written consent of the patient. The results of genetic tests, as part of an individual's medical history, usually are similarly protected.

Despite these and other protections, many people worry that privacy—defined as the right to control access to information about oneself—is rapidly disappearing, especially in highly computerized societies such as our own. In fact, many of the most fundamental questions that individuals who are interested in the HGP ask about the collection and storage of genetic information actually have to do with the larger question of an individual's right to privacy with respect to all types of personal information. The privacy issues that the HGP raises, then, are not new, although now they may be perceived as more serious and more pressing than they have been in the past.

We discuss many of the ethical and public-policy issues that arise from the creation and use of registry databases in Section III. To provide a broad context within which to consider these issues, we describe below some of the general ways in which registries are used in the United States, as well as some of the specific ways in which scientists and health-care professionals use databases containing medical and genetic information. We conclude Section II with a description of several types of databanks containing biological specimens from which both medical and genetic data can be derived.

Registry databases in the United States. Some estimates place the total number of computerized records that describe specific American citizens at more than five billion. Similar estimates suggest that personal information about each of us is moved from one computer to another on an average of about five times each day. In fact, the scope and magnitude of the data collection efforts that already operate in the country prompted the United States Office of Technology Assessment in

1992 to note: "It is virtually impossible for most citizens to know where files about them exist and nearly impossible for individuals to learn about, let alone seek redress for, misuse of their records."

Registries maintained by government agencies. Probably the largest repositories of information about people in the United States are the some 200 federal agencies that collectively maintain nearly 2,000 databases, many with millions of files. These agencies include such huge operations as the Internal Revenue Service, the Social Security Administration, the Immigration and Naturalization Service, and, of course, the Federal Bureau of Investigation. The FBI's National Crime Information Center alone currently holds files on more than 20 million Americans, including not only convicted drug dealers and murderers, but also missing people and people who were arrested but not convicted. The center responds to more than a million inquiries per day at a rate of about 12 per second.

Registries maintained by business and industry. Next to the federal government, the largest repositories of information about specific Americans are credit bureaus. Estimates indicate that considered collectively, credit agencies hold more than 450 million records on more than 160 million people. Depending on the credit bureau involved, these records can contain a variety of information, such as birth dates, social security numbers, mortgage records, current and previous addresses, telephone numbers, employment and salary histories, descriptions of credit transactions of all types, current balances due, information on legal entanglements, family composition, bankruptcies, and tax liens. In fact, the trend in recent years has been to include more, not less, information in such files. Largely in response to the commercial demand for more detailed information about people, many credit bureaus have expanded their information collection policies to include data purchased from banks, retail sales companies, motor vehicle agencies, magazine subscription services, and many other sources. A spokesperson for one of the country's largest credit bureaus recently stated that it buys all the data that it legally can buy. Its stated goal is to compile the most detailed description of every American that can be assembled. Estimates by individuals in the industry indicate that more than 500,000 credit reports are sold each day.

In addition to credit bureaus, other types of businesses also create and maintain registries. Many companies create registries for internal use, such as registries that contain employees' pay, attendance, and performance records, and registries that contain information on customers' orders. Other companies create registries for the explicit purpose of reselling the information. Registries that are created to hold information that will be resold may contain such data as peoples' birth dates, retrieved from the files of state motor vehicle departments (these dates are sold to companies that wish to send their customers birthday cards) and the names and addresses of people who dial particular 800 numbers (these lists are sold to companies that market specific products and services).

Still other companies—probably, by far, the largest category—create registries both to use in conducting their own businesses and to sell. In fact, reselling lists of customers' names, addresses, telephone numbers, and buying habits has become a major industry in the United States. Today, more than 15,000 consumer lists are available for rent, containing more than two billion names. According to one report, these lists include a surprisingly diverse set of categories, from lists of millionaires and the people who live near them, to lists of epileptics, lists of people who are fat or short or thin or tall, lists of compulsive gamblers, and lists of people who just bought a home, just had a baby, just got divorced, or just lost a loved one.

Computer matching of information in registry databases. People who use the data stored in computer-based registries can increase enormously their potential usefulness by a practice called *computer matching*. This practice, developed in the mid-1970s, compares the data held in one electronic registry with the data contained in another. For example, the government has used computer matches of the federal payroll against welfare lists to identify double-dippers, individuals who were collecting both a federal paycheck and welfare payments. Similarly, to focus their marketing efforts more specifically on people who would be interested in and also be able to buy a particular product, businesses often match lists of people who have shown certain buying patterns against lists of people in particular income ranges or with good credit ratings.

Computer matching is a powerful technique for assembling a wide range of information about people into consolidated files, but the process also can reveal things about people that they never meant to make public. In fact, concerns about this practice were part of the impetus behind the Privacy Act of 1974, which prohibits the federal government from conducting computer matches of the data held by one agency with the data held by another. The privacy protections imposed by this act, however, do not apply to the private sector, and computer matching is a widespread practice in business and industry.

Registry databases containing medical information. Registry databases maintained in doctors' offices, hospitals, and by insurance companies and state and federal agencies contain a variety of medical information about millions of people. This information, generally considered to be confidential and protected from casual release by informed-consent procedures, can include personal and family medical histories, laboratory test results, diagnostic information, treatment information, and all of the informal notes and comments that physicians, nurses, and other health-care professionals have added to an individual's file. Because of its potential relevance to a current or former health condition, such databases also may include information about an individual's life style (for example, smoking, eating, and drinking habits), sexual orientation, and mental state.

Health-care workers use registry databases in a variety of ways. Hospitals use computerized registries to store information about their patients so that it is readily available to help health-care workers provide timely and appropriate care. Individual doctors, dentists, and other private practitioners use registry databases to store patients' medical records and to store insurance and billing information.

Many public organizations also use registry databases that contain medical information. For example, data about patients are collected and used by the Medicare and Medicaid programs, by the office of vital statistics in each state, by surveillance programs focusing on birth defects, and by cancer registries. Often, these data are of critical importance to individuals at state and federal levels who must plan for health services in particular regions or for particular constituencies. State departments of health use data derived from medical records to plan for future spending on conditions that may be increasing or decreasing in frequency in that state. Careful study of the data stored in such registries sometimes has suggested that certain conditions are frequent enough to consider screening for those in whom early treatment can make a difference. For example, all babies born in the United States currently are tested for phenylketonuria (PKU), a genetic condition that, if untreated, can lead to severe mental retardation. Early detection allows physicians to prescribe special diets for such babies, thereby minimizing the otherwise serious consequences of the disorder. Other examples of disorders for which newborn infants in the United States are screened include hypothyroidism, galactosemia, congenital adrenal hyperplasia, and hemoglobin abnormalities such as sickle cell anemia and thalassemia.

There are cases, however, in which medical data are used for purposes other than to provide direct health care or to track the incidence of various types of health conditions. For example, life and health insurance companies maintain their own client registries and use medical data from other registries, such as those associated with hospitals and doctors' offices, to identify certain risks in individuals whom they are about to insure. This information can affect the way in which the companies award or set rates for insurance policies. The data in such insurance-based registries are accumulated in a number of ways. Typically, an individual applying for life insurance is asked to complete an application form outlining his or her medical history. The applicant then is required to sign another form granting the insurance company permission to obtain whatever medical files the insurance company deems important to confirm the information provided on the application and to determine insurability. If the face value of the policy is moderately high, the insurance company also may ask the applicant to submit to a range of basic medical tests, such as tests for high blood pressure, elevated cholesterol levels, diabetes, and heart irregularities.

Most insurance companies share the information that they collect about a person with other insurance companies through an organization known as the Medical Information Bureau (MIB). MIB is the largest repository of medical records in the United States, holding information about the health conditions of more than 15 million individuals. Originally created to detect and prevent attempted insurance fraud, MIB now acts as a clearinghouse for medical information. Approximately 750 insurance companies report information about individuals' medical claims and conditions to the MIB, which combines the information into summary health reports and makes these reports available to all of its members. MIB rules prohibit its members from basing their underwriting decisions on MIB derived data, but it is not clear that this policy is enforceable.

One of the concerns that individuals familiar with the handling of medical records in the United States often express is that once medical records are released from a doctor's office, there are no federal laws in place to protect their continued confidentiality. As described above, for example, to file an application for a life insurance policy with one company, an individual generally is required to release his or her full medical history to as many as 750 other companies. It even is possible that employers can obtain detailed medical data without a person's informed consent. Many businesses, including more than half of the Fortune 500 companies, pay for their employees' health care directly out of corporate accounts, even though they may use an insurance company to handle the paperwork. These businesses regularly receive reports on the medical claims that their employees file. Some insurance companies even include the promise in their advertising that they will provide businesses with reports that describe their employees' claim histories. Likewise, if an employee misses work as a result of an illness or injury, an employer has the right to ask for the employee's medical records to verify the cause of the absence and the person's ability to return to work. Because most medical providers will not take the time to purge a file of information unrelated to the illness or condition at issue, all records usually are sent, including information not related to job performance or to that specific absence.

Registry databases and genetic information. To consider the storage of genetic data in registry databases, we must distinguish carefully the genetic information in a registry database created for the purpose of research from the genetic information in a registry database created simply to hold people's medical records. As an example of the former, consider the genetic information stored in a registry database that researchers have developed to help them analyze the inheritance pattern of a particular genetic condition. This information typically is protected very carefully, with tight restrictions not only on who may access it, but also on how it may be used.

In contrast, the genetic data that may appear in an individual's medical records are less carefully protected. Because these data are part of a larger file, they inevitably travel with that file, and are, theoretically at least, as open to inspection as is the health record itself. As we discuss in Section III, this connection between genetic data and health information leads to serious concerns about the privacy and use of information about an individual's genetic status.

Many people have suggested that genetic data are sufficiently different to warrant special protection, as data separate from and potentially more damaging to an individual than standard medical information. The arguments on which this suggestion is based include the observation that one's genetic status is beyond one's control and the argument that people should not be punished (for example, denied health insurance coverage) for such involuntary conditions. People who advocate special protection for genetic data also point out that in the past, genetic data have been used to direct horrifying eugenics efforts (for example, those in Nazi Germany) and that genetic information about one person has implications for other people whose privacy may be violated when a health record is opened. Unlike most standard medical information, genetic data also can be powerful predictors of future risk.

It is not clear, however, that an effort to distinguish genetic data from the larger arena of medical data can be successful. As a practical matter, genetic information can be obtained in many ways, including by inference from an individual's clinical symptoms or family history. Although neither insurers nor employers currently make substantial use of direct results from genetic tests, life insurance providers do use related information such as the family histories that applicants provide on their application forms and the results of cholesterol, diabetes, and blood pressure testing that applicants sometimes must submit to before their application is considered. The OTA report cited earlier also points out that MIB records contain information about several genetic diseases as well as information about "family" diseases. As discussed in Section III, the connection between genetic information and medical information raises the concern that as more powerful testing methods become available, individuals may be faced with choices between obtaining information that could help them make good decisions about their lives and not obtaining that information at all for fear that it would seriously affect their ability to obtain adequate life or health-care insurance or, even, their ability to remain employed.

Databanks. Databases that contain biological materials from which descriptive or identifying information can be derived (for example, samples of blood, saliva, body tissues, or even DNA) often are called *databanks*. Many databanks already exist in the United States and others currently are being developed. Perhaps the oldest such databanks are state storage facilities for PKU blood samples. Others include various forensic databanks that store blood, DNA, or saliva samples from convicted felons; the United States Army and the Defense Department's sample storage programs; databanks associated with private and public genetic research projects; databanks maintained by the Red Cross and other blood donor referral programs; databanks maintained by hospitals, laboratories, and transplant services; and private sperm, ovum, embryo, and tissue banks.

Databanks are created for a variety of purposes. Some databanks are associated with standard screening programs. For example, the Guthrie test for PKU involves placing a small sample of blood on a cardboard disc. A set of such cardboard discs, if labeled and stored with the names of the persons from whom the samples were obtained, is a databank.

Similarly, various government agencies maintain databanks to assist in identification efforts. DNA samples taken from individuals in the armed forces, for example, can help military officials identify the remains of bodies when other means are not available. One way in which officials use such databanks is to compare the pattern of fragments that is generated by restriction endonuclease digestion of the DNA of an unidentified body (its DNA fingerprint), to the pattern generated by equivalent digestion of the stored DNA of soldiers known to be missing in the area. The first extensive use of military DNA profiles occurred during the Persian Gulf War. In early 1996, two United States Marines refused to give blood samples for DNA banking, citing concerns about the confidentiality of the information and potential abuses. Following a series of hearings, the two marines were dishonorably discharged.

Some states have started to collect blood and saliva samples from convicted felons, as a way to help law enforcement officers identify individuals who may have been involved in violent crimes. Sometimes the collection procedures apply both to people convicted of violent crimes and those convicted of so-called nonviolent crimes such as burglary or tax fraud. Comparisons of the fingerprint of the DNA extracted from blood, sperm, or other bodily fluids left at the scene of a crime to a suspect's DNA fingerprint or even to the fingerprints of a range of known offenders can allow law enforcement agencies not only to *exclude* an individual from suspicion (this has long been possible using ABO blood type comparisons), but also to *identify* an individual as a possible or likely source of the DNA involved. Once generated, an individual's DNA fingerprint also can be stored in a registry database, to be examined again and again, as the need should arise. Similarly, once collected, the original biological sample also can be kept to be examined again, in this or other ways.

Health-care organizations also create and maintain specific databanks for diagnostic or treatment purposes. Hospitals, laboratories, blood banks, and transplant services store computerized characterizations of genetic traits such as blood groups, histocompatibility antigens, and serological reactions. Such organizations also often store actual tissue samples collected from individuals, such as blood or kidney tissue, together with the coded information.

Finally, medical researchers, geneticists, and epidemiologists sometimes create or seek access to large DNA databanks to search for clues about the genetic basis of various human diseases. An example of a large registry and databank that scientists use extensively in genetic research is the Centre d'Etude du Polymorphisme Humain (CEPH), an international collaborative effort to promote the mapping of inherited human diseases. CEPH is located in Paris, France, and is funded privately. As a functioning databank, CEPH provides collaborating scientists with actual DNA samples from a large group of cooperating families for which CEPH also has extensive pedigree information. These families serve as a reference for building linkage maps. The use of a common set of families in many different linkage studies is equivalent to using the same strain of mice in many whole-animal studies. These linkage studies establish the positions of markers that scientists can use to identify the location of genetic traits that are inherited in other families. The CEPH databank has been fundamental to the mapping and sequencing objectives of the HGP: in December 1993, for example, the scientific journal *Nature* carried a gene map, produced by CEPH, that contained 2,000 genetic markers covering 87 percent of the human genome.

Databanks that store biological material actually store a great deal of personal information about an individual, including much that we are not yet able to analyze. Because DNA is a very stable chemical, samples taken today for one purpose still could be used in the future for purposes never imagined by the person providing the sample. Methodological advances have made it possible to extract the DNA from even tiny samples of preserved blood or other body tissues. These advances mean that even after years of storage, just about any human biological artifact is a potential repository of personal genetic information. For example, the blood sample preserved on the cardboard of a Guthrie test contains enough biological material for hundreds of additional genetic tests. Note as well the recent interest in testing DNA extracted from samples of Abraham Lincoln's hair, bone chips, and bloodstains to determine whether Lincoln suffered from Marfan syndrome.

Even graveyards contain a wealth of preserved information. In the extension activity associated with this module, students consider this type of genetic information as they analyze the base sequences of several samples of mitochondrial DNA that were taken from bones found in a common grave in Russia.

Because the samples in databanks contain data that potentially are more revealing even than the data contained in many medical records, some health professionals and policy experts have suggested that databanks should be placed under very special types of regulation. These individuals base their arguments on the observation that a DNA molecule is not only a record of what currently exists, but, in some sense, also is a "future diary," though a probabilistic one. It is true that this diary is written in a code that we will understand only in a limited sense for many years. Nevertheless, as we slowly decipher the code—that is, as researchers move forward in their efforts to identify and describe all human genes—those who hold biological samples will be able to learn more and more about any individual whose sample they store, as well as more and more about his or her family. Predictions that might be made on the basis of such analysis would not be precise: we know, for example, that the expression of genetic characteristics varies from individual to individual. Still, as an article in the *Journal of the American Medical Association* (November 1993) points out, "This is information individuals, physicians, insurance companies, employers, and others will want and on which they will base decisions affecting the individual." Such concerns have prompted the American Society of Human Genetics to propose guidelines to minimize the prospects for abuse of the genetic information that one can derive from stored DNA samples.

The Future of Genomic Databases

Genomic databases already play an important role in the HGP. Most people recognize, however, that the structures of today's databases will not be adequate to meet the constantly growing and changing needs that are created by a growing and changing knowledge base. Because we cannot predict exactly what the fully annotated human sequence will look like or how we will use it to ask and answer questions, we cannot predict exactly what the software and hardware required to house it should look like. Even now, however, biological, medical, and computer scientists are trying to envision the final genomic database—its structure, the information it will contain, and the mechanisms that will be available to help people access that information. Such thinking is vital because it affects the choices we make today about database design and use and also because it suggests the types of research in information management that we should pursue to move us toward the database of tomorrow.

Improved design. If electronic databases are to store enormous amounts of information and enable scientists to ask complicated questions and get the answers within a few seconds, they must be designed carefully, according to the types of data that will be in them and the ways in which scientists will use those data. Individuals involved in database design must, therefore, anticipate how the data and their uses will change, and they must structure databases so that they can incorporate new knowledge relatively easily. Even well-designed databases must be evaluated regularly with respect to new knowledge and must be modified periodically to accommodate a rapidly changing science.

Improved capacity for data analysis. To use the information stored in genomic databases, scientists must be able to access and manipulate it. As the volume and complexity of genomic data increase, problems of data analysis also will grow. For example, when a scientist discovers a new sequence, he or she normally compares it with every sequence in GenBank to determine whether the fragment has been sequenced before. Many of the search algorithms in use today require on the order of a few hours to accomplish a full GenBank comparison against a sequence of only 1,000 bases. Unless the individuals involved find ways to speed up the search process, the time required for such a search will increase as the volume of the data stored,in sequence databases increases.

In anticipation of such problems, investigators already are creating new and faster algorithms for accomplishing a variety of data- and computation-intensive processes involving genetic information. Most of these issues, however, will be addressed by external changes completely unrelated to the HGP. The market forces already at work to improve computer speed and data storage densities are enormous. Computers will get faster, cheaper, and have higher storage capacities each year, with or without the HGP. Database projects in science involving weather prediction, studies on global change, and earth-observing satellites already generate data in volumes of terabytes (1012 bytes) per day. The largest genomic databases fall in the few gigabyte range (109 bytes) and have been 10 years in the making. The rate at which genomic data are accumulating, however, is increasing, and the amount of data that must be stored will grow rapidly across time.

A critical computational issue is how scientists integrate several genetic maps to form one comprehensive picture of a chromosome. At this point, largely because of differences in the units of measurement involved, most research groups maintain separate maps for linkage data, physical mapping data, and sequence data. This separation is reflected in the national databases as well, with sequence data located in one type of database and genetic-linkage and physical data stored in another. As researchers generate more data of each type, the pressure to integrate different views of human chromosomes will intensify. In fact, an intermediate goal of the HGP is to develop an atlas that contains one integrated map of each human chromosome. A number of research groups already are at work writing and testing increasingly sophisticated software that is capable of reconciling and displaying simultaneously many different types of genomic data.

Standardization across databases. A controversial issue in database design and information transfer has to do with how the same types of data are stored and represented in different genomic databases. In the early years of genomic research, database development proceeded in rather *ad hoc* ways. As individuals and groups perceived various needs for data management, they quite naturally pursued various ways to meet those needs. The result has been the development of database systems that do not organize or manage data in equivalent ways. This lack of standardization has both technologic aspects (for example, the computer language or equipment used in one database system may be incompatible with those in another) and conceptual aspects (for example, the labels used to refer to particular categories of information may be different).

Significant incompatibilities among different databases make finding and comparing data frustrating and time consuming. When the answer to a user's question is spread across several different databases and each database uses a different format, different definitions for terms, and different access modes, retrieval of the required information sometimes can cost more time and effort than was invested to generate the information in the first place. Incompatibilities among databases also interfere with rapid and easy data transfer, especially transfer from the lab where the data originated to the large public databases. These incompatibilities can become a serious problem, because data that are not accessible are not useful.

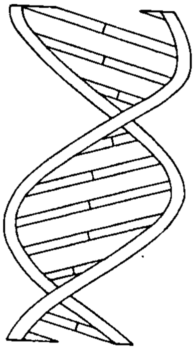
In light of these problems, the biggest initiative in informatics research related to the HGP is likely to be the creation of a federated genomic database, an umbrella architecture that would allow users to search for genomic information without having to worry about where that information resides. For the most part, biologists do not know or want to know which database stores a specific piece of information. They simply want to retrieve the information easily and efficiently, following a particular line of inquiry. This suggests that an important research goal is the development of systems that will make many diverse and independently maintained databases appear to users as a *single database* with a single query language. Scientists refer to such systems as being "seamless." The creation of many seamless gateways would, in theory, allow a user to move from searching GenBank in Bethesda, to searching GDB in Baltimore, to searching DDBJ in Japan, without even noticing the transition. In effect, such gateways would create one functional genomic database from the many that we have now.

Realizing the Benefits

One of the continuing challenges that faces individuals who are associated with the HGP is to ensure that the benefits of ongoing research reach the broader scientific, medical, and industrial communities. Many critics complain that the HGP is taking funding and attention away from other aspects of biological science and that other research projects stand to gain little or nothing from the investment. As the HGP moves forward, those involved in the work, and especially those charged with the responsibility for giving the work its shape and direction, must evaluate constantly what immediate products and tools the HGP can develop for the broader scientific community without jeopardizing its central focus: to describe completely the human genome.

NOTES:

1. Data from GDB (Genome Database).
2. To access these addresses, you must have a web browser; e-mail alone is not sufficient.
3. Cooper, N.G. (ed.) (1992). *Los Alamos Science #20.* Los Alamos, NM: Los Alamos National Laboratory.



Section III Ethical and Public-Policy Dimensions of Research Databases and Registries

There are a variety of potential uses for the information that is stored in research databases and genetic registries, and these uses will affect us in a variety of ways. We can predict some of these ways reliably at present. Others, however, we cannot foresee.

Many people, including scientists, attorneys, specialists in ethics, policy experts, elected officials, executives of insurance companies and of large and small corporations, and citizens have expressed concerns about access to and use of the information in research databases and registries. Two questions capture these concerns:

□ Who should have access to and use of information in research databases and registries?

□ Should society regulate access to and use of information in research databases and registries?

Individuals, institutions (schools, businesses, and other organizations), and citizens will have to deal with these questions, and inevitably, the answers will result in some interests being advanced and others being impaired. Whose interests should receive priority when everyone's interests cannot be advanced and when some interests can be advanced only at the expense of others? This question properly is addressed by ethics and public policy. Because such questions already have arisen and will continue to arise regarding research databases and registries, it is essential that the public understand the ethical and public-policy dimensions involved in answering them. Activities 4 and 5 raise these questions for classroom discussion.

This section begins by describing a number of ethical and public-policy issues that are related to research databases and registries and by identifying for each issue some major ethical and public-policy questions. By considering these questions, we set the stage for more detailed discussions of ethics and public policy and of the most effective ways to address them in the classroom.

Ethical and Public-Policy Questions Concerning Access to and Use of Information in Research Databases and Registries

Geneticists use genomic data that have been collected from many people to do two things. The first is to define the general or "generalized" human genome (aggregate genomic data) in terms of such items as catalogs of genes, details of the human map, and lists of DNA and protein sequences. Databases that store this type of generic information are known as research databases. GDB, GenBank, and OMIM are examples of research databases.

A second way in which geneticists use genomic data is in the creation of registries, which are databases that contain genomic data about specific individuals (personal genomic data). It is common practice today that an individual's consent is secured before personal genomic data are entered into a registry database. Registries may record genetic conditions or susceptibilities of individuals or even may contain sufficient data that one can identify individuals on the basis of their genetic profiles. Scientists use genomic registries to follow the inheritance of particular traits or DNA sequences from one family member to another (for example, in genetic linkage analysis) or to investigate possible genetic contributions to a variety of conditions (for example, hypertension).

Geneticists also use such data to help them assess the need for genetic services in an area. In the Great Plains Genetic Service Network, for example, non-identifiable personal data are collected and reported in eight states on people who seek genetic services of various types. This collection effort depends on voluntary participation by the individuals involved. While it is the case that the collected data do not reflect all of the people in the region who seek genetic services, the information nonetheless is useful for understanding what services are provided to whom and for justifying state programs that support genetics services.

Issues concerning research databases. Table 4 lists some of the ethical and public-policy issues related to research databases. This section elaborates those issues.

Table 4 Some issues concerning research databases.

- cost of research databases
- access to research databases by those who did not pay for them
- use of research databases to define "normality"
- use of a generalized genome in a multiracial society and world
- potential dangers of counseling oneself

Cost of research databases. The estimated cost of the United States's share of the HGP will be more than $150 million per year for a period of approximately 15 years. Some critics have argued that at this level of funding, the HGP takes scarce economic and human resources away from research on "small science" projects, especially those that are targeted to specific, well-known diseases and the individuals affected by those diseases. Critics say that the needs of individuals with these diseases are urgent and pressing, whereas the needs served by the HGP are less pressing and more remote. On the other hand, those who support the HGP point out that it already is producing valuable new knowledge about the structure and function of genes, about the role of genes in evolution, and about biological variation, as well as aiding in the discovery of disease genes. They also point out that although $150 million is a lot of money, it is small compared with that being spent on AIDS research or on many NASA projects. Supporters argue that new knowledge that we gain from the HGP has the potential to affect the health of millions of Americans in the very near future.

There also are considerable costs associated with maintaining research databases. Many users of research databases such as OMIM use the data for private purposes, for example, for the care of patients, for which they earn money. This amounts to an indirect public subsidy to individuals such as physicians and to institutions such as hospitals for the purposes of private benefit. Questions about the cost of the HGP and related research databases include the following:

□ Should taxpayers support the large science project of creating research databases at the expense of potentially valuable, small science projects?

□ Should those who make money from the use of publicly funded research databases be charged reasonable user fees to help cover the development and maintenance costs of the databases?

Access to research databases by scientists and others from countries that did not contribute to the cost of producing the databases. The HGP is an international effort, but only a few countries are heavily involved in funding the work. These countries include, for example, the United States, Canada, Japan, the United Kingdom, Germany, France, Spain, Italy, and Australia. Yet those involved with the genome project do not restrict access to research databases only to scientists in contributing countries. Universal access reflects the important scientific values of cooperation and sharing of information. On the other hand, perhaps other countries could afford a modest contribution to the cost of developing and maintaining databases, thus freeing resources in countries such as the United States for other purposes. A key access question about research databases, then, is:

□ Should anyone from any country in the world have access to research databases, regardless of whether that individual's country contributes to the cost of developing and maintaining the databases?

Use of research databases to define "normality." The HGP will provide new insights into human genetic variation. Scientists use DNA sequencing techniques, for example, to study variation among different alleles for various genes and to help them understand how some genetic changes result in phenotypic differences and others do not. Scientists also study variation in the expression of genes in individuals and in populations and, in some cases, are beginning to investigate the molecular basis of variation in severity and age of onset among groups of individuals who have the same genetic disorder.

Although scientists estimate that any two people differ at only about 0.1 percent of their bases (that is, at the molecular level, we are more alike than we are different), this variability, nonetheless, involves about 10^6 differences from one person to another. This variability has important consequences for each of us and is an important and legitimate focus of genomic research. As we accumulate data about human variation in research databases, however, we risk the possibility that individuals or organizations will begin to apply it in simplistic ways. For example, information from research databases could be used—or misused—by individuals, institutions, or governments to define normality and abnormality and to allocate public resources to institutions or individuals on the basis of these definitions. Such an action might create a climate where individuals or populations could be subject to genetic discrimination based solely on genotypic data. Questions related to our growing understanding of human variation include the following:

□ How can data from research databases help us understand the concepts of genetic health and genetic disease?

□ Should data from research databases be used to define normality or abnormality solely on the basis of genotypes?

□ Is genetic discrimination ever ethically justifiable?

Use of a generalized genome in a multiracial society and world. Some critics of the HGP express concern that the generalized human genome that is described in research databases is drawn, to a considerable degree, from samples taken from Caucasian Americans and Europeans. As a consequence, databases will contain information about human genetics that may not reflect the racial diversity of *Homo sapiens*. Some research supported by the HGP addresses this concern, although the HGP has no formal policy on this matter at present. Questions about the ethnic mix reflected in genomic data include the following:

□ Should research databases contain information on a generalized human genome that reflects the racial diversity of our species?

□ If so, should the HGP have a formal policy to ensure that the generalized genome reflects the racial diversity of the human species worldwide?

Potential dangers of counseling oneself. Some individuals may access research databases for the purpose of counseling themselves about risks, prevention, and treatment of genetic disorders. However, these databases do not contain data about the probability of genetic disorders in any particular individual. Moreover, data in research databases usually are quite sophisticated, and can be incomplete and even erroneous. These facts call into question whether counseling oneself would lead to informed decision making. At the same time, individuals who are not health-care professionals already have access to complicated, possibly incomplete and erroneous information in medical books and journals in public libraries and in libraries in hospitals, medical schools, and major medical centers. One could, therefore, argue that access to data in research databases poses no additional ethical problems. Questions about allowing unrestricted public access to research databases include the following:

□ Should those who manage research databases restrict access to these data if the data may be used for the purposes of counseling oneself about the risks, prevention, and treatment of genetic disorders?

Issues concerning registries. Ethical and public-policy issues about genomic registries include issues of informed consent, control of access to personal data, and restrictions on the use of personal data. These issues are not unique to genomic registries. For example, they also apply to registries containing credit information, to registries at hospitals and in doctors' offices that contain general medical information, and even to registries that store information about personal buying practices. It is possible, however, that our ever-increasing capacity to generate and to use genetic data may make us more aware of these issues and may force us to deal with them sooner than we might have otherwise. Table 5 summarizes some of the ethical and public-policy issues that genetic registries raise.

Informed consent. It is routine in medical practice to obtain the informed consent of any patient for diagnostic and therapeutic interventions. It also is routine to obtain the patient's consent to release information for research purposes or to health insurance companies that pay for the patient's care. Informed consent requires the health-care provider to explain the medical situation to the patient. The informed patient then makes a decision to authorize (informed consent) or to refuse to authorize (informed refusal) either medical intervention or release of information or both. The policy of informed consent gives individuals the opportunity to make decisions about their own health and to protect the privacy of information about themselves.

Table 5 Some issues concerning genetic registries.

- informed consent
- control of access to personal genomic data
- restrictions on the use of personal genomic data

The bioethics literature reflects strong consensus among ethicists, public-policy experts, and health-care professionals that informed consent always should be required before obtaining genetic material for analysis and entry into a registry. Some individuals, however, take the view that there also are no exceptions to the requirement for informed consent for access to or release of information about patients. In contrast, others take the view that there are reasonable exceptions, particularly if such access or release could protect other innocent people from threats to their health or well-being. Questions relating to informed consent include the following:

□ Should information about a person's genetic profile be disclosed to that person without his or her consent? Such a question might arise in the case of a research subject who gives blood for one purpose, but is found as a result of testing to have a condition unrelated to the research project. Although the new information might benefit the person, its disclosure may not have been covered by the original consent agreement. Thus, such a discovery raises the question of whether the new information should be disclosed to the individual involved.

□ Should access to an individual's personal genomic data be allowed to benefit other persons without that individual's informed consent? Such benefit might occur when blood or tissue is tested from an ill or deceased family member. If the testing reveals the presence of a potentially harmful allele, should that information be shared with other family members who might benefit from the knowledge?

□ Should researchers be allowed to access and use personal genomic data for community-based projects, for example, to identify the prevalence of a genetic trait or disease among a particular racial group? A related question has to do with whether research scientists should be allowed to obtain information from DNA labs that offer clinical services in linkage analysis or direct DNA testing.

Control of access to personal genomic data. There are other ethical and public-policy issues concerning access to personal genomic data that are independent of questions of informed consent. These issues concern the obligations of those who manage databases to protect the data from unauthorized access. This concern is very real for electronic records because computers containing registries likely will be linked in large area networks (including international networks), just as some research databases now are linked. Questions relating to unauthorized access to personal genomic data include the following:

□ Who among the employees of a registry should be provided routine, authorized access to personal genomic data that are identified by name?

□ Should government agents such as public-health officials or law enforcement officials such as judges be allowed access to personal genomic data that are identified by name?

□ Should private agencies such as employers or health-insurance companies have access to personal genomic data that are identified by name?

□ Should there be laws that explicitly regulate access to personal genomic data that are contained in a registry or elsewhere? Should individuals be charged a fee for access to their own genomic data?

□ Should individuals or married couples be provided access to their own genomic data for the purpose of counseling themselves about genetic risks and options?

Restrictions on the use of personal genomic data. Once access is granted to personal genomic data, separate and important questions arise about the permissible uses and applications of such information. A key consideration here is that such information always is incomplete, may contain errors, and is subject to developing scientific and clinical interpretations. For example, so-called junk DNA may at some future date be found to play a significant role in the expression of some human traits. Questions about the future uses of genomic data include the following:

□ Should educators be allowed to use personal genomic data to identify high school students for particular academic programs, for example, honors, sports, or other special educational tracks?

□ Should personal genomic data be used to select the daily diets of children or as a basis for providing or withholding health care?

□ Should a school board authorize faculty access to student records to identify students at risk for learning disabilities such as those associated with fragile X syndrome?

Ethics

The foregoing questions concern what society should or should not do in response to the issues that research databases and genetic registries raise. What one should do or should not do defines one's duties. A duty or obligation is something that an individual, institution, or society is bound to do because it is the right thing to do, and as such, duty justifiably restricts freedom. To understand the ethical and public-policy dimensions of research databases and registries, individuals must address the following questions:

□ What is ethics?

□ What is public policy?

□ How are ethics and public policy related?

Activities 4 and 5 are designed to address these questions in the classroom. Every student will confront these issues directly or indirectly for the rest of his or her life, and the intellectual skills that students will develop during these activities are essential to becoming responsible citizens in the face of the ethical and public-policy challenges of research databases and registries.

Ethics is the study of what is right and wrong and what is good and bad, applied to the actions and character of individuals, institutions, and society.

Although ethics often is considered an esoteric, difficult discipline that has little practical value, you and your students make ethical judgments all of the time. Ethics as an intellectual discipline involves analysis of and arguments for the justifications for our decisions.

Generally, in ethics we apply the terms *right* and *good* to those actions and qualities that foster the interests of individuals, institutions, and societies. We apply the terms *wrong* and *bad* to those actions and qualities that impair the interests of individuals, institutions, or society. During the last 2,500 years, Western philosophers have developed a variety of powerful methods and a reliable set of concepts and technical terms for studying ethics.

The features of ethics. Experts generally agree on the following features of ethics and the teaching of ethics. First, ethics is a process of rational inquiry, and inquiry involves posing questions. Rational inquiry in ethics, as in science, involves posing clearly formulated questions and seeking *well-reasoned* answers to those questions. These activities require us to analyze and clarify key terms and concepts. Well-reasoned answers to ethical questions constitute arguments. Ethical analysis and argument, then, result from successful ethical inquiry.

Second, ethics requires a solid foundation of information and rigorous interpretation of that information. For example, to ask and answer questions about the ethics of research databases and registries, we must have a solid understanding of genetic science (in particular, a strong understanding of the complex and variable interactions of genotypes and environments that produce phenotypes), as well as a good grasp of computer and database technology. Thus, ethics is not strictly a theoretical or abstract enterprise, but is concerned in a vital way with practical matters, such as science and technology.

Third, because trade-offs among interests are complex, constantly changing, and sometimes uncertain, there often are competing, well-reasoned answers to questions about what is right and wrong or good and bad. This is especially true of complex matters such as access to and use of information in research databases and registries. Genetic variation and its variable expression in individuals are fundamental features of the human genome. Because this variation makes possible competing interpretations of the data in research databases and registries and because individuals have different value systems, the most frequent outcome of ethical inquiry about research databases and registries will be competing, well-reasoned answers about permissible access and use. A major pedagogical goal of Activity 4 is to have students develop and reflect on such answers to ethical questions about genetic registries.

How to talk about interests. The concepts and terminology of ethics that are relevant to research databases and registries involve different ways of talking about the interests of individuals, institutions, and society (Table 6). You likely will find that your students use many of these concepts and terms as they discuss the issues posed in Activities 4 and 5.

Interests are stakes that individuals, institutions, and society have in the outcomes of decisions and events.[4] There are two ways to talk about these

Table 6 A taxonomy of moral considerations.

Consequences	Human Relationships	Rights	Justice
results	care	claims	substantive (outcome)
outcomes	sentiments	entitlement	procedural (process)
effects	feelings	power	
purposes	suffering	liberty	
goals		autonomy	
ends		freedom	

interests. One way is to talk in terms of the *consequences* of actions. Synonyms the students might use include *results, outcomes, effects, purposes, goals,* and *ends*. Rational discussion about consequences requires that individuals be able to give well-founded reasons to explain why society should or should not pursue the consequences of an action. Consequences that advance interests are labeled "right" or "good," and society should pursue such consequences. An example from research databases is the potential for increased knowledge of genetic contributions to disease, which then would advance the public-health interests of society. Consequences that impair interests are labeled "wrong" or "bad," and society should not pursue such consequences. An example from research databases would be to define the generalized human genome in a way that is insensitive to the multiracial character of the world's population. This would impair the interests of all because it may promote racial stereotypes.

The "ethics of care" is a recent and relevant development in the ethics literature. The concept of the ethics of care is based on empirical research on moral psychology (how people make moral judgments and decisions). The research indicates that the quality of human relationships is affected by the moral judgments that individuals make about the consequences of various actions and decisions.

Obviously, genetic information has consequences for the quality of human relationships. Indeed, sometimes genetic information can be interpreted properly only in a family context. Therefore, you should expect students to express their moral concerns and judgments along these lines. For example, the knowledge that a spouse may carry a potentially harmful dominant allele can affect the marriage relationship and the relationship of that adult with his or her children, adult siblings, and parents. The concern for relationships also can extend beyond family to one's community. The effect on communities may become an important dimension to consider as society considers questions about access to and use of information in research databases and registries. The ethics of care underscores the importance of the consequences of decisions that may affect the quality of human relationships. Terms students might use to express this idea include *care, sentiments, feelings,* and *suffering.*

A second way to talk about interests is in terms of a *right* or *rights*. A right is a claim that one should be treated in a certain way regardless of the consequences. Synonyms the students might use include *claim, entitlement, power, liberty, autonomy,* and *freedom*. An example of a right that is pertinent to this module would be an individual's informed refusal to provide genetic material for laboratory analysis and inclusion in a databank.

Respect for rights promotes individual interests because doing so allows individuals to pursue things that individuals value, even if other people in the society do not value those things. Society ought to respect and protect individual rights. Denial of rights damages individual interests because doing so does not allow individuals to pursue what they value. To release genetic information about an individual to a health insurance company when that individual explicitly has denied such access would violate that individual's right.

Denial of any right should, as a rule, be prevented. Yet, denial of some rights may be justified when individuals' rights are in conflict or when the consequences of exercising a right would result in serious, irreversible harm to others. Rational discussion about rights requires that individuals be able to give credible reasons to explain why society should respect or should deny the right in question.

Interests pertain not only to individuals, but also to groups of individuals and to institutions. The management of conflicting interests at this level in ethical analysis is considered in the context of justice. Justice, broadly stated, requires that we render to each individual or institution what is due to him or her or it. There are two ways in which "due" can be understood. The first is termed *substantive justice,* which involves what the *outcome* of a decision-making process should be. The second is termed *procedural justice,* which involves the *process* by which decisions about outcomes should be reached.

There are competing theories of substantive justice in Western philosophy. Some take the view that a just outcome meets basic needs such as needs for food and shelter. Others take the view that a just outcome is based on merit, that is, on what individuals have earned by working and using their talents. Still others take the view that outcomes should, as far as possible, be an equal distribution of resources. And some argue that different principles of distribution apply to different spheres of human life.

This debate about substantive justice is centuries old, and philosophers have been unable to reach consensus on *the* theory of substantive justice. As a result, you may expect your students to express a variety of views on substantive justice during Activity 5. How can one respond effectively as a teacher to this variety of interpretations? The answer to this question is found in the considerable consensus that exists around the view that decisions should be made in a way that meets a basic requirement of *procedural justice*. This is the requirement that everyone whose interests are potentially at stake in a decision should have his or her interests taken into account. Thus, for example, any decision about access to information in a registry would require, as a matter of procedural justice, taking into account the interests of all of the individuals in the registry. One way of doing so is informed consent.

The role of argument in ethical inquiry. Ethics establishes what should or should not be done on the basis of rational argument. An argument is a set of clearly stated premises or reasons that together *justify* a conclusion. All credible arguments must meet two standards.

The first standard is *validity*. An argument meets this standard when the conclusion follows from the reasons by the accepted rules of logic. The classroom activities for this module are designed so that validity is satisfied in all cases. The principal advantage of this design is that you do not have to teach the formal rules of logical inference to your students.

The second standard requires that *all premises or reasons count as good reasons for everyone*. To satisfy this standard, a student must be able to say why everyone should accept a particular reason as important and relevant to the issue at hand. It is not enough for a student to say that the reason is important to him or her. Just as scientific evidence must be public, the reasons for ethical arguments also must be public. The classroom activities in this module place a great emphasis on meeting this standard.

Reasons that are supported by religious arguments deserve special consideration. Students always should respect their classmates' religious beliefs and convictions. However, the moral content of religious beliefs and convictions frequently is expressed in terms and concepts that are important only to the person making the argument or to one's particular faith community. An essential teaching strategy, then, is to have students try to express a particular religious belief or conviction in terms that are important to everyone. If this process succeeds, the reason counts as important to all. If this process does not succeed, the reason does not count as important to all and cannot be used in an argument. This result *in no way* affects its importance for the student who offered it or the obligation of the other students to acknowledge and respect that importance. Denigration of any religious belief is inconsistent with respect for students and you should not tolerate it in the classroom. Teachers should take this same approach with all expressions of serious moral convictions, including those of nonreligious origin.

Recurrent ethical concepts related to research databases and registries. You may expect students to raise a number of ethical concepts that are related to research databases and registries in classroom discussions. Concepts related to registries include privacy, confidentiality, autonomy, paternalism, and discrimination (Table 7).

Table 7 Recurrent ethical concepts in teaching ethics of registries.

- privacy
- confidentiality
- autonomy
- paternalism
- discrimination

Privacy refers to the right to control access to information about oneself. Typically, privacy is thought to be among the most basic rights of citizens in a free society. Each of us has the right to control access to information about ourself. Privacy of data is not a major issue for research databases because these databases contain only generalized genetic information. In contrast, privacy is a major issue for the creation and use of registries. Registries do contain data about individual genomes that can be identified by name. Students will confront issues of privacy in each of the activities in the module.

Confidentiality is an obligation of those who obtain information about individuals (legitimately or illegitimately) to protect the privacy of that information. The obligation of confidentiality in health care, which is a relevant context for a registry, is based on both consequences and rights. Maintaining confidentiality usually results in individuals being more willing to share sensitive information about themselves with health-care professionals—information that is essential to diagnose and manage health concerns effectively and safely. Maintaining confidentiality also is a form of respect for the right of privacy of individuals.

These two ethical bases combine to create a very strong obligation to individuals to maintain confidentiality. However, the obligation does have limits. For example, it may be ethically justified to break confidentiality if that is the only way to protect the lives of innocent others. Some state courts have adopted this view with respect to psychiatric patients who express detailed homicidal plans about identifiable individuals.

A challenge for your students will be to develop arguments about limits on the right to privacy and the obligation of confidentiality with respect to registry information. This is a difficult challenge because the lives of innocent others usually are not at stake in the privacy and confidentiality of a registry. The obligation of confidentiality, therefore, will be harder to break.

Autonomy is a centerpiece of Western political philosophy. Autonomy usually is understood to mean individual self-determination. An individual should not be made subject to the rule and control of others unless there is sufficient ethical justification for doing so (for example, to protect others' rights). Autonomy often is defended on the grounds that free individuals know best how to define and pursue their own interests.

Autonomy also is defended on the grounds that respect for autonomy produces highly valued consequences for all. When people are left alone to pursue their interests, society is better off. It is important to recognize, however, that because genetic information is family information, one person's autonomy may conflict with another's. For example, several members of a family may choose to participate in some type of registry, while another relative chooses not to participate. In this case, information might be inferred about the nonparticipating relative from the participants' information. Depending on one's interpretation, one might say that the nonparticipating relative's choice was overridden or violated by the participation of other family members.

Paternalism is the counterpoint to autonomy. Paternalism involves the claim that someone else can know an individual's interests better than that individual and also can know better how to pursue those interests. The word "paternalism" has its roots in the Latin word for "father," connoting the traditional claim of parents to know better than their children what is in the best interests of those children. Obviously, paternalism is an affront to autonomy. Students who wish to justify paternalistic strategies for regulating access to and use of information in genetic registries must show why respect for autonomy is justifiably limited in those circumstances. For example, one might argue that access to a registry for the purpose of unsupervised self-counseling should not be allowed on the paternalistic grounds that too many individuals could be confused or overwhelmed by genetic information. These people might, therefore, be at risk for making choices about diagnostic and therapeutic interventions that are not in their own best interests.

Discrimination involves making choices on the basis of differences between things. Discrimination that is ethically justified appeals to differences that are shown to be ethically relevant considerations, considerations that promote valued consequences, respect for rights, or substantive or procedural justice. Denying a driver's license to someone who is totally blind is an example of ethically justified discrimination because adequate eyesight is relevant to the safe operation of a motor vehicle. Discrimination that is ethically unjustified appeals to ethically irrelevant differences. Differences between things or people or institutions are irrelevant when no one can find any basis for them in well-made ethical arguments. Denying a driver's license to someone because he is a male is an example of unjustified discrimination because there is no scientific or other reliable evidence that being a male undermines or destroys the capacity to operate a motor vehicle in terms of consequences, rights, justice, or any other ethical concept.

Discrimination is not a central ethical issue in this module, although students may raise it as a concern. The challenge for them will be to show on the basis of ethical argument that a genetic difference indeed should count as a basis for ethically justified discrimination.

Public Policy

Public policy is a set of guidelines or rules that results from the actions or lack of actions of governmental entities. Governmental entities act by making laws. Laws can be made by each of the three branches of government in the American political system: by legislatures (statutory law), by courts (common law), and by regulatory agencies (regulatory law). All three types of law are pertinent to research databases and registries. The law will be concerned for the most part with the regulation of access to and use of information in research databases and any local, state, regional, or national registries that might be created. When public policy is a function of law, it is called *de jure* (according to law) public policy. When no laws exist to regulate behavior, public policy is called *de facto* (actual) public policy.

Students often do not appreciate the large role that regulatory law plays in science and technology. Regulatory law is written by the executive branch of the government, under authorization by the legislative branch. We can expect regulatory law related to research databases and registries to grow steadily during the coming years. Such laws will be promulgated primarily at the federal and state levels, with local regulatory law playing a limited role. In the absence of law to the contrary, private institutions will make *de facto* public policy to guide and regulate institutional behavior on the basis of their own approach to ethical analysis and argument.

Law relates to substantive and procedural justice in complex ways. One useful way to understand this complex relationship is the following. On matters as challenging as questions about access to and use of information from genomic databases, there inevitably will be competing accounts among citizens and institutions about which goals and outcomes society should pursue as a matter of public policy based on substantive justice. In writing law, policy makers will, in effect, adopt or endorse—explicitly or implicitly—one or more goals and outcomes, while rejecting others.

However, the goals and outcomes that policy makers reject will not necessarily lose their ethical justification on the basis of substantive justice. Instead, these rejected alternatives will continue to be raised and debated among citizens and in policy-making branches of government, as part of the ongoing ethical inquiry into *de jure* and *de facto* public policy. This means that any particular policy likely will not be permanent. That is, policy makers and the public will have to revisit public-policy decisions periodically. It is important for your students to understand that ongoing ethical inquiry and debate is one effect of the competition among different accounts of substantive justice.

Such debate, which is a central feature of American political life generally, reminds us that procedural justice plays an important role in public-policy decision making. When matters are unsettled, it is important that the ongoing ethical inquiry and public debate take into account the interests of all affected individuals, communities, and institutions. In fact, one effective strategy of procedural justice in a pluralistic society such as that of the United States is to rely for a time on *de facto* public policy as a form of social experimentation. In time, the best results of such *de facto* policy then become candidates for *de jure* public policy. Thus, the public-policy challenges of research databases and registries fit well within our political traditions, and we should not assume necessarily that these challenges will overwhelm our capacity as self-governing people to respond effectively to them.

Legal Protections of Privacy of Genetic Information in the United States

At present in the United States, genetic information is treated like medical information. There currently is no existing federal privacy law about the data contained in DNA databanks or registries, although such laws have been proposed. However, some states have adopted laws and rules concerning access to and use of genetic information. These laws vary and, for the most part, place restrictions on the insurance industry.

At the time of this writing, twelve states (Alabama, Arizona, California, Colorado, Florida, Louisiana, Maryland, Montana, North Carolina, Ohio, Tennessee, and Wisconsin) prohibit the use of genetic information by insurers. Most of the states' restrictions narrowly apply to only one or two conditions or to carriers of a genetic disorder who remain unaffected by it. Two states (Arizona and Montana) broaden restrictions to include all single-gene and chromosomal conditions.

Except in Wisconsin, Ohio, and Colorado, all of these statutes allow insurers to use genetic information that can be supported by actuarial data or, in some cases, reasonably anticipated experience. In 1992, Wisconsin passed a law that prohibits health insurers from requiring or requesting individuals (a) to take a DNA test, (b) to reveal whether they have undergone a DNA test, or (c) to disclose DNA test results. The statute also prohibits insurers from using DNA test results to determine rates and other aspects of health insurance coverage. Ohio's law (passed in 1993) calls for a 10-year moratorium on the use of information from genetic tests. Colorado's law was passed in 1994 and prohibits the health, group disability, and long-term care insurance industry from seeking, using, or keeping genetic testing information for underwriting or nontherapeutic purposes. A special committee of the HGP is charged with proposing guidelines to regulate the growing use of genetic testing.[5]

Discussion regarding the privacy of medical information dates back at least to 1974. The Privacy Act of 1974 established among other things, the Privacy Protection Study Commission (instead of a permanent regulatory commission). The Commission's 1977 report remains the most authoritative statement on large record-keeping systems in the United States. Some recommendations of this report include (a) greater federal and state legislation pertaining to medical records; (b) state statutes that recognize individual rights of access to, correction of, and confidentiality of medical records; and (c) rules restricting the disclosure of information to that level necessary to accomplish the purposes for which the disclosure is made. These recommendations have, in general, not been adopted.

With the recent discussion about health-care reform, there has been renewed interest in the privacy of medical information. President Clinton's Health Care Reform Plan acknowledged that "currently, no uniform, comprehensive privacy standards related to health-care information exist" and that it is important to consider such standards.[6]

Many people believe that it is time to think critically about the specific rules that govern the collection, storage, and distribution of genetic information. These people argue that this is the case not only because of the increased availability of genetic information but also because of some unique characteristics that it possesses. Such characteristics include the fact that (a) one's genetic make-up is beyond one's control; (b) genetic risk assessments are assumed to have more predictive power than other types of health risk assessments; (c) genetic information has implications for other people, particularly those in one's immediate and future family; (d) genetic risk information can be psychologically and socially stigmatizing; and (e) DNA is a very stable chemical and samples taken for one purpose may be used at a later date for other, possibly unauthorized, reasons.[7]

Some individuals support federal as opposed to state legislation for DNA databanks and registries, arguing that uniform state legislation is difficult to achieve. However, even federal legislation is considered problematic by some because many databanks and registries are in private hands and thus cannot be governed by federal legislation. An attractive alternative for some people includes voluntary agreement. For example, the genetics networks that are established across this country could contribute to the development of voluntary standards for DNA databanks and registries.

Interest in regulating genetic information is fueled by a commitment in the United States to privacy. But what is privacy? According to George Annas, there are four major senses of privacy.[8] The first three describe aspects of the constitutional right to privacy. The central sense of privacy, found in the liberty interests that are protected by the Fourteenth Amendment, is the right to privacy that forms the basis for the opinions by the United States Supreme Court that limit state interference with intimate, individual decisions, such as those involving reproductive decisions. The second and third senses of constitutional privacy protect certain relationships (such as husband and wife, and health-care professional and patient) and certain places (such as the home) from governmental intrusion. The fourth sense of privacy, the common-law right to privacy, applies to private actions: "the right to be let alone" (for example, the right to keep personal information inaccessible to others).

Three (1, 2, and 4) of the four senses of privacy apply to genetic information. First, genetic disease differs from contagious disease insofar as genetic disease is controlled through reproductive decisions and actions as opposed to pharmaceutical agents. Second, as the activities in this module illustrate, genetic information is family information that often is "discovered" in the relationship between the health-care professional and the patient. Third, genetic information is personal and unique in that it affects self-identity. Empirical evidence shows that the knowledge or the assumption that one carries certain disease genes can affect self-perception.

This overview suggests that we can expect legislation related to research databases and registries to increase steadily during the coming years. As our knowledge of and ability to manipulate the human genome increase, we will be confronted with growing concerns about what society ought to permit. The role of public policy is to address such concerns and to offer a forum for resolving them.

Teaching Ethics and Public Policy in the Classroom

Ethics is vital to *de facto* public policy because it provides the concepts and terminology for the type of carefully organized public debate that can result in well-reasoned conclusions about what society should or should not do as a matter of law. This inquiry is valuable in and of itself. After society identifies a well-reasoned conclusion, it is reasonable to ask whether it should be enacted into *de jure* public policy. Sometimes the best public policy response is *not* to enact a law in response to a controversy, but rather to allow individuals, institutions, and society to act in the manner they choose.

Ethical analysis and argument, on the one hand, and public policy, on the other, therefore are often well integrated. This is one of the underlying assumptions of this module and of Activities 4 and 5. Another underlying assumption is that public policy can anticipate and prevent ethical problems. Thus, public policy, when it is informed by ethical reflection based on good science, is a powerful form of preventive ethics.

The teaching materials for this module are organized according to the following five steps for teaching ethics and public policy in the classroom. Steps 1-4 involve teaching ethics as a process of rational inquiry; Step 5 involves teaching the students how they can translate the conclusions of ethics into public policy.

□ Step 1—*information gathering*. Science is the key to this step because genetics, molecular biology, and technology provide ethics with information about research databases and registries and their scientifically valid applications. Activities 1-3 explore various aspects of this step.

□ Step 2—*evaluation of the information gathered in Step 1.* This step involves the evaluation of information about genomic databases and the application of that information in terms of the impact of genetic data on the interests of individuals, institutions, and society. Activities 4 and 5 involve this step. For example, in Activity 4, students think in terms of both rights and consequences as they make arguments in response to the PKU case. In Activity 5, the students move from ethical analysis and argument to making public policy related to a range of cases involving genetic registries.

□ Step 3—*making ethical arguments.* As noted above, an argument is a set of clearly stated reasons that together justify a conclusion. An argument begins with reasons and ends with conclusions. Activity 4 emphasizes the need for students to justify the premises or reasons in their arguments by stating why those reasons should count as important to everyone. Expect your students to offer a variety of reasons to meet this standard. Also expect the students to support the same conclusions with different reasons. Obviously, different reasons can lead to different conclusions.

□ Step 4—*analyzing arguments.* The students should compare their arguments and attempt to produce the most well-reasoned argument. Activity 4 is designed to promote this process.

□ Step 5—*translating the results of ethical arguments into public policy.* Activity 5 helps the students with this step. Your students should see that making public policy is challenging, partly because policy attempts to address differing, if not competing, ethical concerns. Students also may see that whether a well-reasoned ethical argument should become *de jure* public policy depends on whether that conclusion satisfies certain basic requirements. For example, listen for your students to comment on such matters as the costs of implementing particular policies (including possible financial, social, or personal costs), the urgency of implementing a new policy, how effective a particular policy is likely to be, and whether appropriate means exist to implement the policy.

Conclusion

Our students will live their adult lives in an era that will be increasingly influenced by the HGP, and it is important that they have some idea of its objectives, scope, assumptions, and potential implications. It is our intent that the activities included in this module will provide your students with the knowledge and skills required to understand research databases and registries at a basic level and to analyze the relevant issues in a manner that fosters informed, respectful debate and sound decision making.

NOTES:

[4] Feinberg, J. (1984). *Harm to others.* New York: Oxford University Press.

[5] Task Force on Genetic Testing. (1996). *Interim principles.* NIH-DOE Working Group on Ethical, Legal, and Social Implications of Human Genome Research.

[6] The White House Domestic Policy Council. (1993, draft version). *The president's security plan: The Clinton blueprint.* New York: Times Books.

[7] Brown, R. S. and Marshall, K. (eds.). (1993). *Advances in genetic information: A guide for state policy makers.* 2nd edition. Kentucky: The Council of State Governments.

[8] Annas, G. (1993, November). Privacy rules for DNA databanks. *Journal of the American Medical Association* 270(19): 2346-2350.

Implementation Support

Your students can complete the seven activities included in this module in about two weeks, depending on the amount of time you allot for computer interaction and class discussion.

Conceptual Organization of the Activities

We have organized the first six activities to form a conceptual whole that moves students from an introduction to genomic databases (Introductory Activity), to an examination of their structure and function (Activities 1-3), and then to discussions of the ethical and public-policy issues that the use of such databases raises (Activities 4-5). The final activity in the module (Extension Activity) draws the students' attention back to the scientific use of genomic data in the study of evolutionary relationships. Because the extension activity deals with a specific example of the scientific use of genomic data, you may wish to complete it as part of your work with this module. Alternatively, you may choose to complete it independently, in conjunction with your students' study of evolution.

Introductory Activity: ***The HGP and Electronic Databases*** **(30 minutes).** This activity introduces students to the Human Genome Project, a coordinated research effort that will generate more data than any other single effort in the history of biology. To record the complete sequence of the haploid human genome would require the equivalent of 200 telephone books of 1,000 pages each. Variation in the sequences among humans will increase further the amount of data that must be stored.

Completing this activity helps students understand that the data collection, storage, and distribution requirements of the HGP are such that scientists must use electronic databases. Students begin by performing three manual searches of a short DNA sequence (2,508 bases) to identify specific target sequences. Students record the number of target sequences that they locate in each one-minute search and then repeat their searches on the computer using an identical computer-based sequence and an electronic search algorithm. Students compare the results of the two types of searches and discover that the computer-based retrieval of sequence data is much more rapid and accurate than the manual search techniques.

Activity 1: ***Genetic Registries*** **(2 class periods).** This activity introduces students to the two types of genomic databases used in the HGP (research databases and genetic registries) and offers students an opportunity to explore the Local Genome Database (LGD), which is the model registry that we provide with the software included in the module. Students take on the role of specific individuals in a set of three fictitious families whose genetic profiles are stored in the LGD. They learn that registry databases contain information that is identified as relating to specific individuals by moving among the various files in the LGD to gather and analyze specific health and genetic information about "themselves" and their fictitious families. Teams of students use the information that they gather about family relationships to construct a pedigree for their fictitious family. To complete these pedigrees, students explore the personal genomic data that are contained within the LGD and begin to discover for themselves some of the reasons that access to information in registry databases typically is restricted to authorized individuals.

The activity concludes by asking the students to decide whether their fictitious person's DNA should be tested for the presence of three additional alleles (the alleles associated with cystic fibrosis, sickle cell disease, and hypertrophic familial cardiomyopathy). Students decide in Activity 4 how to handle the new data that are gathered through the testing that they authorize.

Activity 2: ***Explaining the Outliers*** **(2 class periods).** Students verify the pedigrees that they constructed in Activity 1 against those stored in the LGD and discover that the database contains data that appear to be contradictory: there is one individual in each of the three families who is not included in the computer-generated pedigree. Teams of students use information in both the LGD and the National Genome Database (NGD), which is the model research database included in the software, to develop and test a set of hypotheses about the reasons for these discrepancies.

This activity provides students with an opportunity to explore a research database and to reinforce their understanding of the differences between a research database and a genetic registry. As the teams collect information to help them test their hypotheses, the students also consider several issues of importance to the HGP, such as the relationship between genotype and phenotype, the importance of allelic variation that is not associated with human disease, and the fact that databases can contain errors of various types.

Activity 3: ***Genetic Anticipation*** **(1 class period).** As students complete Activity 3, they learn that HGP-related research is yielding a wealth of answers to many questions of immediate interest, and they encounter some of the personal and societal issues that are raised by this rapid increase in our knowledge of the human genome. Specifically, the students consider the case of a young woman who is struggling to understand the implications of her genotype for fragile X syndrome, one of several disorders that HGP investigators have been able to associate with a newly discovered disease mechanism, the expansion of areas of trinucleotide repeats during transmission of the DNA from one generation to the next.

Students learn that this young woman discovered her genotype for this gene when she innocently investigated her personal genetic information in the LGD and then cross-referenced it to the information in the NGD. During the course of the activity, the students follow the path of her research, examine the issues and the options that she faces, and consider some of the questions that arise when the pace of genetic research and the availability of genetic information outstrip an individual's readiness to deal with the new information.

Activity 4: ***Who Should Control Information about My Genes?*** **(1 class period).** In Activity 4, the conceptual focus of the instruction changes from an emphasis on the science and informatics of the HGP to an emphasis on the ethical issues that HGP-related research raises. One of the questions that the young woman in Activity 3 faced was to what extent she should keep her genetic data private and to what extent she should share it with other members of her family and community. In Activity 4, the students examine this same question, this time on behalf of the fictitious people whom they represent and with respect to the new genetic data that were collected by the DNA testing that they authorized in Activity 1.

Using the skills of ethical reasoning, students analyze ethical issues related to the release of these data. They consider several possible courses of action: to disclose only anonymous, combined data related to the testing; to disclose personal data, but only to the specific individuals involved; and to announce the personal data publicly. The activity concludes with the students deciding whether the new data should be entered into the LGD.

Activity 5: ***Making Public Policy*** **(1 class period).** In this activity, students work as elected legislators to recommend a federal policy related to the protection of genetic information that is stored in genetic registries and DNA databanks. Class discussion builds on the understandings that students gained during Activity 4 about ethical issues related to genomic information and also reflects the scientific and technical knowledge that students acquired about research databases and registries during Activities 1-3. Students begin to see that each policy option advances some interests over others and that the act of debating and making public policy is a way that a democratic society promotes the discussion of complex issues and governs itself in the face of competing interests and values.

Extension Activity: ***HGP Data and Evolutionary Biology*** **(2 class periods).** The extension activity demonstrates how scientists use sequence data to study both family and evolutionary relationships. In the first part of the activity, students use the database to evaluate the similarities and differences among the mitochondrial DNA (mtDNA) of nine skeletons recovered from a shallow grave in 1991 and one living person. Using a 12-base sequence from mtDNA, the students develop and test a hypothesis about the relationships among these ten individuals, and learn that we can use DNA sequence comparisons to identify relationships between individuals of the same species.

In the second part of the activity, the students use the database to study DNA and amino acid sequences from several different species. Based on their knowledge of the species and on the results of sequence comparisons, students construct proposed evolutionary trees for these species, and learn that we can use DNA and amino acid sequence differences to identify evolutionary relationships between different species.

Preparing to Teach the Activities in the Classroom

Before you begin teaching the activities in this module, you will need to load the software onto the computers that you will be using and to duplicate the required student text materials and worksheets. Table 8 summarizes the materials that you will need to have prepared for your class to complete the activities and also lists the relevant software settings and passwords. Refer to the Advance Preparation section of each activity for specific information about preparation.

Student background. Your students should have a background in basic genetics, including pedigree construction and interpretation, before they begin the activities in this module. They also should be familiar with the molecular structure of DNA and with the relationship between the base sequence of the DNA and the amino acid sequence of proteins specified by that DNA. Students should be familiar with the words in the glossary marked with an asterisk (*).

The computer-based activities do not presume more than a very basic understanding of computer manipulation. If students already are familiar with the operation of a mouse, they will spend less time learning to use one. The mouse manipulations are explained thoroughly and students quickly will become comfortable with them. Except in those cases that are specifically identified in the teacher's annotations, the software and the student directions for using it are completely self-contained and self-explanatory.

You may wish to invite guest speakers into your classroom to discuss some of the disorders mentioned in the activities. You also may wish to introduce your students to ethics and public policy several days before you begin the module. Section III of the teacher narrative contains suggestions about teaching these topics in the classroom.

Organizing groups. We have designed all of the activities to be completed by groups of students working together. On the basis of extensive field testing, we recommend that you organize your groups according to the guidelines provided in Activity 1. These guidelines will help you construct groups in sizes that are appropriate for the sizes of the extended families that the students study in Activities 1 and 2. Students in groups that are larger than the recommended sizes may become overwhelmed by the extra data that they collect. Likewise, even if your student-to-computer ratio is one-to-one, students still will need to share the data that they retrieve with other members of a group in order to complete their pedigrees.

Although the number of students that work together in each group is less critical for the other activities, we recommend that you keep your students in the same groups for all of the activities in the module. This will allow each group to develop a shared experience with the software and with the fictitious family that it encounters in Activity 1. This shared experience also will enhance your students' perceptions of the activities as a conceptual whole. This will be particularly important in Activities 4 and 5, as students consider the ethical and public-policy implications of the data that they have been retrieving from the databases.

If your student-to-computer ratio is greater than six students to one computer, you will need to change the way you teach the module from what we describe in the activities. For example, if you have only one computer available, you may want to have students complete the computer-based work across an extended time period. You can do this in several ways. The most practical way is to

Activity	Estimated Time	Materials Required	Database Setting(s)	Password(s)
Introductory Activity: *The HGP and Electronic Databases*	30 minutes	• student text, p. S-1 • Worksheets 1, 2, and 3 (Worksheet 3 is assigned as homework preceding Activity 1.) • overhead transparency of BLM T-1	*The HGP and Electronic Databases*	Watson
Activity 1: *Genetic Registries*	2 class periods	• student text, pp. S-7–S-9 • Worksheets 4, 5, 6, and 7 (Worksheet 7 is assigned as homework preceding Activity 2.) • copy of BLM T-2 • overhead transparency of BLM T-3 (optional) • cotton-tipped swabs • small envelopes	*Genetic Registries*	Lyon
Activity 2: *Explaining the Outliers*	2 class periods	• student text, pp. S-17–S-18 • Worksheet 8	*Explaining the Outliers, Part A*	Sutton
			Explaining the Outliers, Part B	Avery
Activity 3: *Genetic Anticipation*	1 class period	• student text, pp. S-25–S-26 • Worksheet 9	*Genetic Anticipation*	McClintock
Activity 4: *Who Should Control Information about My Genes?*	1 class period	• student text, pp. S-29–S-30 • Worksheet 10 (Worksheet 10 is assigned as homework preceding Activity 5.) • envelopes with swabs (from Activity 1) • overhead transparencies of BLM T-4, BLM T-5, and BLM T-6 • 9" x 12" envelope • butcher block paper, markers, transparent tape (optional)	*Genetic Anticipation**	McClintock
Activity 5: *Making Public Policy*	1 class period	• student text, pp. S-35–S-36 • Worksheets 11, 12, and 13 • one copy of Worksheet 14 • overhead transparency of BLM T-7 • name tags (optional)	*Genetic Anticipation**	McClintock
Extension Activity: *HGP Data and Evolutionary Biology*	2 class periods	• student text, pp. S-43–S-46 • Worksheets 15, 16, 17, and 18 (Worksheet 16 is assigned as homework preceding Part B.) • overhead transparencies of BLM T-8 and BLM T-9	*HGP Data and Evolutionary Biology, Part A*	Mendel
			HGP Data and Evolutionary Biology, Part B	Lederberg

Table 8 Summary of the materials and database settings for the activities. *Note that there are no special database settings for Activities 4 and 5. The setting for Activity 3 will make all of the data in the LGD and the NGD available for your students' reference.

have groups of students come in before or after school, beginning at least two weeks before you plan to teach the module. This will allow the students time to retrieve the information that they need from the computer and to begin the assignments. In this case, you will need to delay the discussions until all students have completed the relevant work on the computer.

Another strategy is to use your computer as a center along with several other centers at which students complete other activities. In this strategy, students would rotate through the computer center, eventually completing all of the activities.

A third way to structure the activities if you only have one computer available is to give the students printouts of selected data and ask them to work from those printouts to complete their computer research. This final strategy, however, will not give the students an opportunity to discover how the databases are organized. It also will not help them learn to search the databases independently. You can add some of this experience by using an overhead projection system to demonstrate searching the databases. Or, you might allow individual students to demonstrate searching the databases. Use this strategy only if you have no other options.

The physical setting. We suggest that you arrange the computers so that they are separated by enough space that students in separate groups will not interfere with each other. The students also will need room to move around when they are exchanging positions.

When the students are discussing issues within their groups, they will need a suitable working space. During full-class discussions, the students also will need to get away from the computers. When students have the opportunity either to discuss an issue or to play on a computer, they almost always will choose to play. Because of this tendency to continue to work on the computer while discussions take place, you will find that moving the students away from the computers or turning off the screens will help increase particpation.

Encouraging student interaction with the software. The activities are designed to allow all of your students to work at the computer keyboard. Some students are better typists than others and may attempt to help their classmates by doing their keyboard entry for them. We encourage you to move throughout the classroom to make sure that every student has a chance to work with the databases. As discussed below, knowledge about a computer is not as important as knowledge about how to use computers. Each student should have the opportunity to experience, through direct interaction with the software, the power of electronic databases and the importance of accurate work at the keyboard.

Dealing with Values and Controversial Issues[9]

Instructors sometimes feel that the discussion of value issues is not appropriate in the science classroom or that it detracts from the learning of "real" science. This module, however, is based upon the conviction that there is much to be gained by involving students in analyzing issues of science, technology, and society. Society expects all of your students to function as citizens in the democratic process, and their school experience should provide opportunities for them to learn how to deal with contentious issues with civility, objectivity, and fairness. Likewise, students need to learn that science affects life in many ways. Opportunities to consider some of these ways also will reinforce those scientific principles that we desire to teach.

The activities in this module provide opportunities for the students to discuss, interpret, and evaluate human genetic research in light of values and ethics. Many issues that students will encounter —especially the privacy of genetic information and related public-policy questions—are potentially controversial. How much controversy develops will depend on many factors, such as the degree of homogeneity that exists among your students with respect to socioeconomic status, perspectives, value systems, and religious preferences. It also will depend on how you handle your role as facilitator of the discussion. Your language and attitude may be the factors that are most important to the flow of ideas and the quality of exchange among the students.

Neutrality is probably the single most important characteristic of a successful discussion facilitator. The following behaviors will help you guide your students in discussions in which factual information is balanced with feelings.

□ Encourage your students to discover as much information about the issue as possible, and ask questions that will help your students distinguish between those components of an idea or issue that scientific research can answer and those components that are a matter of values. Students should understand the importance of accurate information to any discussion and should recognize the importance of distinguishing factual information from opinions.

□ Keep the discussion relevant and moving forward by questioning or posing appropriate problems or hypothetical situations. Invite your students to respond to or to build on each other's ideas. Avoid asking questions that have exact answers unless the facts are important to the integrity of the discussion. Encourage everyone to contribute, but do not force reluctant students into the discussion.

□ Use unbiased questioning to help the students critically examine all views presented. Help your students consider different points of view thoroughly by asking them to define the relevant arguments and counterarguments. Let the students help you promote the expression of alternative points of view.

□ Allow for the discussion of all feelings and opinions. Avoid becoming a censor of views that are radical or shocking (as long as these views are consistent with the facts). When a student seems to be saying something for its shock value, look to see whether other students recognize the inappropriate comment and invite them to respond.

□ Avoid seeking consensus on all issues. The multifaceted issues that the students discuss result in the presentation of divergent views, and students should learn that this is acceptable. In some cases, however, helping the group reach consensus on a compromise solution to a problem may demonstrate compromise as a powerful determinant of cooperative community action.

□ Keep your own views out of the discussion. Experts in science education recommend that teachers withhold their personal opinions from students. The position of teacher carries with it an authority that might influence students. The danger also exists that the discussion might slip into indoctrination into a particular value position, rather than an exploration of divergent positions. Either result misses the point of the activities. If your students ask what you think, you may wish to respond with a statement such as "My personal opinion is not important here. We want to consider your views."

□ Acknowledge all contributions in the same evenhanded manner. If the class senses that you favor one group of ideas over another, you will inhibit open debate and discussion. For example, avoid praising the substance of contributions. Instead, praise the willingness of students to contribute by making such comments as "Thanks for that idea" or "Thanks for those comments." As you display an open attitude, a similarly accepting climate will begin to develop within the class.

□ Emphasize that everyone must be open to hearing and considering diverse views. Point out that we cannot make intelligent decisions if we close ourselves off from some viewpoints. Even if we cannot agree with or are offended by a viewpoint, we still must hear it so that we know that it exists and can consider it as we shape our own views.

□ Create a sense of freedom in the classroom. Remind students, however, that freedom implies the responsibility to exercise that freedom in ways that generate positive results for all. If necessary, remind them as well that there is a fine line between freedom and license. In general, freedom is a positive influence, while license usually generates negative results.

□ Insist upon a nonhostile environment in the classroom. Do not allow your students to make *ad hominem* arguments (arguments that attack the person instead of the idea). Help your students learn to respond to ideas instead of to the individuals presenting those ideas.

- □ Respect silence. Reflective discussions often are slow. If you break the silence, your students may allow you to dominate the discussion.
- □ Finally, at the end of the discussion, ask your students to summarize the points that they and their classmates have made. Let your students know that your respect for them does not depend on their opinion about some controversial issue. If students feel that they must respond in particular ways to gain your approval, your class will not discuss issues openly and with forthrightness.

Following these general suggestions should help you stimulate meaningful student-to-student interaction with as little direct involvement by you as possible. Initially, some students may have difficulty responding without specific direction. It is important, however, that you resist the temptation to intervene extensively in the initial, sometimes uncomfortable phase of long silences and faltering responses. Unless students are given opportunities to evaluate ideas and values in the context of a larger problem, they may never learn to do so.

Using the BSCS Databases

Like a microscope or a pH meter, a computer is an important tool for a scientist to know how to use. We have designed the activities in this module to help students begin to see the role of computers in the HGP and also to have the opportunity to use the computer as a research tool. The activities do not require students to possess more sophisticated technical skills than the ability to manipulate the mouse and to enter search criteria. Nevertheless, to complete the module, students will have to learn to search a database in order to retrieve information that is relevant to a question or task and will have to learn to analyze and apply that information appropriately. As students begin to develop these skills, they are learning to use the computer as a research tool.

General description of the software. The central section of the software consists of two databases, the National Genome Database (NGD), a model of a research database, and the Local Genome Database (LGD), a model of a registry database. Although the two databases are different, they are organized so that the user can move back and forth between them without seeming to go to a different program.

As explained in the teacher narrative, a database is a collection of information organized into a general format. The first (or most specific) level of organization in an electronic database is called a *field*. In the NGD, a field contains specific information about some aspect of a gene such as its name, map position, or base sequence. In the LGD, a field contains specific information about a fictitious person such as his or her name, age, or sex.

At the next level of organization in an electronic database is the *record*. A record is a collection of fields. In the NGD, a record contains information compiled about a single gene (that is, it contains the information that occurs in all of the fields related to that gene). In the LGD, a record consists of the collected information about one individual.

The National Genome Database. The NGD is a simulated research database that contains information on several genes. It has fields consisting of gene name, symbol, definition, organism, map position, notes on the underlying biology, and allelic sequences. Each record consists of two screens: a *General Information* screen that provides general information about the gene, and an *Allelic Variations* screen that lists and describes the gene's various allelic forms. You can search the NGD on gene name, gene symbol, and sequence.

The Local Genome Database. The LGD is a simulated registry database that contains information about 52 fictitious individuals. Each record includes the individual's name, age, sex, current social status, parents' and siblings' names, personal medical history, family medical history, allelic sequences, family pedigree, and a unique sample number. For each record, you will have access to three screens. The *General Information* screen contains background information about the individual (for example, age, sex, the names of other family members, and personal and family medical histories); the *Genotype* screen contains genotypic information on four genes; and the

Pedigree screen displays the full pedigree of the extended family to which the individual belongs. (As discussed below, the *Pedigree* screen is not visible unless the data for Activity 2 or Activity 3 are activated.) You can search the LGD on name, gene symbol, sample number, sex, and sequence.

Additional data. The Introductory Activity and the Extension Activity require the students to use additional data that are not in either the LGD or the NGD. For example, the Introductory Activity uses a sequence of approximately 2,400 bases that students search for the presence of shorter target sequences. Similarly, in the Extension Activity, the students use selected DNA and amino acid sequences to study family and phylogenetic relationships. The data for the Introductory Activity and the Extension Activity become available to the user when the appropriate software settings are activated.

The dynamic nature of the software. As indicated above, we have designed the software so that both the data and the functions available to the user change from one activity to the next. For example, the data that are associated with Activity 1 do not include the family pedigrees or information about the significance of the allelic sequences for fragile X. In contrast, the data associated with Activity 2 include the pedigrees, and the data associated with Activity 3 include both the pedigrees and information about fragile X. Table 9 summarizes the data and the functions that are available for each activity.

After your class has finished a particular activity, you will need to change the software setting to activate the data and functions for the next activity. You will find a blackline master on p. TS-3 (BLM T-3) that you can use to make an overhead transparency that lists the steps required if you want your students to change the activity settings on their computers.

System requirements. We have designed the software for use on Macintosh computers or on computers using Windows. Table 10 lists the system requirements for Macintosh and Windows machines. The two versions of the software are

Database Setting	Data Available	Functions Available
The HGP and Electronic Databases	unidentified sequence data	sequence search; count function
Genetic Registries	LGD (without pedigrees); NGD (without information about fragile X alleles)	standard search functions for the LGD and NGD
Explaining the Outliers, Part A	LGD (with pedigrees); NGD (without information about fragile X alleles)	standard search functions for the LGD and NGD
Explaining the Outliers, Part B	LGD (with pedigrees); NGD (without information on fragile X alleles)	standard search functions for the LGD and NGD; retest function
Genetic Anticipation	LGD (with pedigrees); NGD (with updated information on fragile X)	standard search functions for the LGD and NGD
HGP and Evolutionary Biology, Part A	mtDNA sequence data	replace matching bases with dots; cluster function
HGP and Evolutionary Biology, Part B	amino acid sequence data	replace matching amino acids with dots; cluster function

Table 9 The dynamic nature of the software.

very similar, but not identical. For example, the Windows software requires that users close a screen before accessing the next screen. The Macintosh software is more transparent; the screens will close and open automatically. If the instructions for use are significantly different between the two platforms, you will find separate instructions for each system. Otherwise, you will see a set of instructions that is appropriate for both systems.

Although the software will run on systems with the characteristics listed in Table 10, the faster the machine, the faster you will receive a search result. Older, slower machines may appear to have crashed during searches. If the pointer turns to a watch or a spinning ball (Macintosh), or to an hourglass (Windows), the computer is executing a search and you should *not* click the mouse again.

Using the software. The best way to learn to use the software is to do the activities as the students would do them. Therefore, we do not provide detailed operating instructions here. Instead, we suggest that you start the software and go through all the activities yourself before you use the software with students. If you are a novice computer user, remember that one of your best sources of information may be a knowledgeable student. If you require technical assistance with the software, please call BSCS (719-531-5550) and ask for technical assistance for the seond genome module.

Installing the software on your computers. Before you can use the software, you must install it on each computer that you will use. The compressed software will fit on a single disk. Before installing, make sure that you have the necessary system software and sufficient space on your hard drive. You will not need additional software for either the Macintosh or the Windows version. Figure 12 lists the steps required to load the software onto a computer and Figure 13 lists the steps required to start the software.

Table 10 System requirements.

	Macintosh	Windows
System	6.05 or later	3.1 or later
Hard drive	1.2 Meg	1 Meg
Free RAM	1.5 Meg	1 Meg

NOTE: The software will not run on a network; you must load the software onto each machine.

Special software features. The software has two special features that may be useful to you and your students as you complete the activities in this module. Because these features have somewhat specialized uses, we do not describe them in the computer instructions that appear in the activities.

□ *Searching for a sequence in the NGD or the LGD.* In the Macintosh version, if you click the mouse on one of the DNA sequences recorded for an individual in the LGD, or on the DNA sequence of an allele in the NGD, you will see the sequence highlight and then return to normal. If you now initiate a search on *Sequence*, you will see the highlighted sequence as the choice in the pop-up box. If you click on *OK*, the computer will retrieve a list of all occurrences of that sequence in that database (the NGD or the LGD).

□ *Printing from the databases.* The Macintosh version of the software provides two options for

Installing the Software

For Macintosh Computers

- Place the BSCS_HGP disk into the disk drive. A screen should appear. If it does not, double-click on the BSCS_HGP disk.
- Double-click on the file *BSCSDB.SEA Installer.* The BSCS logo screen will appear. Click on *CONTINUE.*
- Unless you wish to change the default folder, click on *INSTALL* and the software will extract automatically and be placed in a new folder titled BSCS_HGP.
- Eject the disk by dragging its icon to the trash can.

For Windows Computers

- Start Windows and select the Program Manager.
- Place the BSCS_HGP disk into the appropriate drive and, from the Menu Bar, select *FILE.*
- Select *Run*, type in the appropriate drive letter, then type <:\setup>, e.g., b:\setup. Click on *OK.* The program will begin the installation. Unless you wish to change where the files are placed or edit the default name of the software, click on *OK* at any choice.

Figure 12 Installing the software on a computer.

Starting the Software

For Macintosh Computers

- If your hard drive is not open, double click on the hard drive icon.
- Find the folder named *BSCS_HGP* and double click on it. (If necessary, use the scroll-bar arrow to find this folder.)
- Double click on the file named *BSCS Databases* to start the program.
- The first screen that will appear is the BSCS logo screen. Following that, the *Human Genome Project: Computers, Biology, and Privacy* screen will appear.
- Click and pull on *ACTIVITY* at the top of the screen until the appropriate activity title is highlighted. Enter the password for the activity and click on *OK*.

For Windows Computers

- Start Windows if it is not active.
- Open the application window, *BSCS*, and double click on *BSCS Databases*.
- The first screen that will appear is the BSCS logo screen. Following that, the *Human Genome Project: Computers, Biology, and Privacy* screen will appear.
- Click and pull on *ACTIVITY* at the top of the screen until the appropriate activity title is highlighted. Enter the password for the activity and click on *OK*.

NOTE: Windows software requires that users close a screen before accessing the next screen. You may have to explain this characteristic to your students.

Figure 13 Starting the software.

printing from the database. Both options work in both the NGD and the LGD. The first option is a print screen function; selecting this option directs the printer to print the specific screen that you are viewing. The second option is a print record option. This option allows you to print all of the information in the database about a specific individual or a particular gene.

Figure 14 lists the instructions for using these print options. Only the print screen function is available in the Windows software.

Hints for using the software successfully. The following additional information may help you and your students use the software more successfully.

Printing from the LGD and the NGD

To Print a Screen (Macintosh and Windows)

- Retrieve the screen that you wish to print.
- Click and pull on *FILE* at the top of the screen until *Print Screen* is highlighted (*Print Form* in Windows).
- Release the mouse and the program will print that screen.

NOTE: If you want to use a keystroke combination instead of the mouse, press [command-P] (Macintosh) or [CTRL-P] (Windows) to print the page.

To Print a Record (Macintosh)

- Retrieve the record that you wish to print.
- Click and pull on *FILE* in the menu bar at the top of the page.
- Highlight *Print Report* and release the mouse.

NOTE: If you want to use a keystroke combination instead of the mouse, press [command-P].

Figure 14 Instructions for printing from the software.

□ We have used the conventions listed in Figure 15 in the computer instructions. You may wish to ensure that your students are familiar with these conventions before they begin the activities.

□ One way to reduce the time that you spend searching the databases is to search using criteria that will give you a list. For example, if you are looking for information about several members of the same family, try entering only the last name. In this case, the computer will retrieve a list with the names of all the members of that family. Clicking on a name in the list will take you to that person's file. Clicking on *RETURN TO LIST* will return you to the list. This function allows you to select another person from the list and retrieve his or her record without repeating the earlier search.

□ Likewise, when you are searching on *Name*, you do not need to enter the entire name, but just

< >	Type the letters enclosed in these symbols, but do not type the symbols.
[ ]	Press the key named inside the brackets; e.g., [return] means to press the return key.

Figure 15 Conventions used in the computer instructions.

enough to denote a unique individual or a unique gene. For example, if you enter "Anna Ma" on a name search, you will see the record for Anna Major. On the other hand, if you search on the name "Ann," you will retrieve a list with four names. If you search on "Anna," you will see two names.

□ If you misspell a name, or put in extra characters or spaces, the computer will not locate the individual's file. You may wish to caution your students that a common problem is that users enter an extra space between the first and last names of the person whose record they wish to retrieve. Successful searching requires that users spell accurately and avoid entering extra spaces.

□ Users frequently forget to click *BEGIN SEARCH* to start a search. Warn your students that if the computer does not seem to be searching (that is, if the cursor has not turned into the symbol appropriate to the software), they may have forgotten to initiate the search. Simply clicking on *BEGIN SEARCH* will correct this problem.

□ On some monitors, the pop-up boxes on the *Pedigree* screens may not appear to be matched to the pedigree symbols. If you experience this difficulty, check the font size setting for your computer; font size should be set to normal, not large.

Quitting the software. Figure 16 explains how to quit the software.

Quitting the Software

For Macintosh Computers

- Click and pull down on *FILE*, highlight *Quit*, and release the mouse.
- After the program has quit, click on the folder labeled *BSCS_HGP* and then click again in the close box in the upper left corner to close the folder.

For Windows Computers

- Click on *FILE* on the menu bar, highlight *EXIT*, and release the mouse.
- Close the window by clicking on the close bar.

Figure 16 Instructions for how to quit the software.

NOTES:

9. The section *Dealing with Values and Controversial Issues* has been adapted from *Basic Genetics: A Human Approach*, 2nd edition by BSCS, published by Kendall/Hunt Publishing Company. Copyright ©1990 by BSCS.

Glossary

This list of abbreviations and definitions is intended to help you understand the background information and teacher annotations provided in this module; it should not be used as a test for students. Students should, however, be familiar with the words indicated with an asterisk (*) before they begin the activities. The glossary is divided into three sections: Section I includes terms about genetics, ethics, and public policy; Section II includes computer terms; and Section III includes genetic disorders.

Section I: Genetics, Ethics, and Public Policy

***allele:** an alternative form of a gene; any one of several structural variants of a gene.

ATCC: American Type Culture Collection. ATCC provides researchers with information about possible sources for a variety of specialized reagents, probes, and cell lines that are useful in genetic research. ATCC is maintained in Rockville, Maryland.

autonomy: individual self-determination.

***autosome:** any nuclear chromosome other than the X or Y chromosome.

***base:** a purine (A or G) or a pyrimidine (T or C) on one strand of the DNA that forms hydrogen bonds with a complementary base on the other strand. The number of base pairs sometimes is reported as a measure of the physical length of a segment of DNA.

***base sequence:** the order in which the bases occur in a DNA molecule.

CEPH: Centre d'Etude du Polymorphisme Humain. CEPH is an organization located in Paris, France, that maintains an international database of actual DNA samples from a large group of families for which scientists also have extensive pedigree information.

cloned DNA: fragment of DNA that has been isolated and inserted into an appropriate vector (for example, a plasmid or a yeast artificial chromosome) for replication.

***codon:** a sequence of three nucleotides in DNA or RNA that specifies an animo acid.

confidentiality: the obligation of those who obtain information about individuals (legitimately or illegitimately) to protect the privacy of that information.

conserved sequence: a base sequence (or an amino acid sequence) that has remained relatively unchanged throughout evolution.

contig map: a physical map showing the order in which specific DNA fragments occur adjacent to each other (contiguously) along a chromosome.

***de facto* public policy:** guidelines or rules made by individuals or institutions in the absence of law to the contrary. *De facto* means according to fact, the way things actually are.

***de jure* public policy:** written guidelines or rules made by agencies of government (legislatures, courts, and regulatory agencies of the executive branch). *De jure* means according to law.

discrimination: making choices on the basis of ethically relevant or irrelevant differences between things, events, or agents.

DOE: U.S. Department of Energy.

ELSI: Ethical, Legal, and Social Implications. A working group composed of members from the Department of Energy, the National Institutes of Health, and other organizations. The ELSI working group is funded by both the DOE and the NIH and is charged with coordinating research, discussion, and proposals about the societal implications of the HGP.

ethics: the branch of philosophy that considers questions of good or bad and right or wrong.

exon: a segment of a gene that ultimately is incorporated into mRNA.

fragile site: a gap or defect observed in the continuity of a stained chromosome.

GDB: Genome Database. GDB is a gene mapping research database maintained at Johns Hopkins University in Baltimore, Maryland.

GenBank: a genetic sequence research database maintained at the National Center for Biotechnology Information at the National Library of Medicine in Bethesda, Maryland.

***gene:** the hereditary unit of DNA that occupies a certain spot on a chromosome, has a specific effect on the phenotype, and can mutate to various allelic forms.

gene mapping: the process of determining the location of genes at specific sites on chromosomes.

genetic anticipation: a phenomenon in which the severity of symptoms associated with a genetic disorder increases (and/or the symptoms appear at an earlier age) in successive generations.

genetic counseling: the educational process that assists individuals, couples, or families with decisions related to genetic disorders with which they may be affected, for which they may be at risk, or for which they may be carriers.

genetic linkage map: a map in which the distance between markers relates to the frequency with which the markers are inherited together rather than being separated by recombination during meiosis.

genetic profile: the details of an individual's genome.

genetic screening: the search in a population for individuals having genetic characteristics 'that likely will be harmful to themselves or to their descendants.

genetic testing: the search in an individual for genetic characteristics that likely will be harmful to themselves or to their descendants.

genome: the DNA content of an individual, including all 44 autosomes, the 2 sex chromosomes, and all the mitochondrial DNA. For the HGP, one each of the different chromosomes—22 autosomes, plus X, plus Y, plus a mitochondrial chromosome.

***genotype:** the genetic constitution of an organism.

***heterozygous:** having two alleles that are different for a given gene.

HGP: Human Genome Project.

***homozygous:** having two alleles that are identical for a given gene.

informed consent: a process in which an individual makes a decision based on relevant information.

interests: the stakes that individuals and institutions have in the outcomes of decisions and events.

intron: a segment of a gene that occurs between adjacent exons. An intron is transcribed into nuclear RNA, but is removed during subsequent processing and does not appear in the mRNA.

justice: the rendering to each individual or institution what is due to him, her, or it.

marker: a trait, gene, or fragment of DNA that can be identified on a genetic or physical map.

MEDLINE: a bibliographic database produced by the National Library of Medicine in Bethesda, Maryland. MEDLINE covers the fields of medicine, nursing, dentistry, veterinary medicine, and the preclinical sciences.

methylation: the process by which particular cytosines in a DNA molecule are enzymatically converted to 5-methylcytosine by addition of a methyl group (-CH_3) to carbon 5 in cytosine. About 80 percent of mammalian DNA is methylated, and cellular methylation patterns are inherited. Hypermethylation of cytosine within or near a coding sequence is associated with a reduction in gene activity. Genes that are permanently active in all cells lack methyl groups.

MIM: *Mendelian Inheritance in Man*. MIM is the print version of OMIM (Online *Mendelian Inheritance in Man*).

NIH: National Institutes of Health.

***nucleotide:** a subunit of a nucleic acid. A nucleotide consists of a five-carbon sugar joined to a phosphate group and a nitrogen base.

OMIM: Online *Mendelian Inheritance in Man*.

OMIM is a clinical research database maintained at the Center for Medical Genetics, Johns Hopkins Hospital, Baltimore, Maryland.

open reading frame (ORF): the sequence between two stop codons in the same reading frame. A scientist searching for a gene in a segment of sequenced DNA might look for an open reading frame.

paternalism: the claim that someone else can know an individual's interests better than that individual and also can know how to pursue that individual's interests better than that individual.

***pedigree:** a graphic record of the inheritance of a particular trait through many generations of a family.

penetrance: a term that describes a gene's expression. Individuals carrying a mutant gene with reduced penetrance may or may not manifest clinical symptoms of the related condition.

***phenotype:** observable characteristics of an organism that are produced by the organism's genotype interacting with the environment.

physical map: a map in which the distance between markers is the actual distance. Scientists developing a physical map may report the location of a gene or marker by describing its relationship to a particular chromosomal band, its relationship to specific restriction endonuclease recognition sites, or as the actual number of base pairs from some specified landmark.

PIR: Protein Information Resource. A protein sequence database maintained at the National Biomedical Research Foundation in Washington, D.C.

polymorphism: the existence of two or more genetically determined forms. Biologists use the term to describe genetic variation at many levels, including the phenotypic level (for example, the existence of two or more forms of a particular trait or characteristic) and the genotypic level (for example, the existence of two or more base sequences associated with a particular segment of DNA).

POSSUM: Pictures of Standard Syndromes and Undiagnosed Malformations. POSSUM is a CD-ROM product that provides clinical information about a variety of genetic and congenital disorders.

predisposition: a tendency or inclination toward something in advance.

privacy: the right to control access to information about oneself.

procedural justice: refers to what ought to be the process of making decisions regarding what is due someone or what is due an institution.

public policy: a set of guidelines or rules, both written and unwritten, that results from the actions or the lack of actions of governmental agencies.

reading frame: a sequence of bases, conceptually organized into groups containing three bases each. An "open reading frame" follows an initiation codon and could be translated into an amino acid sequence.

recombination: the natural process of breaking and rejoining DNA strands. Recombination produces new combinations of genes and, thus, generates genetic variation.

restriction endonuclease: an enzyme that recognizes specific base sequences in the DNA and cuts the DNA at defined locations with respect to these sites.

right: a claim to be treated in a certain way regardless of the consequences of doing so.

sequence-tagged site (STS): a short section of DNA found to have a unique base sequence when compared with the rest of a particular genome. Scientists use STSs as important landmarks in the construction of physical maps.

substantive justice: refers to what ought to be the outcome of a decision-making process regarding what is due someone or what is due an institution.

tandem repeat sequences: base sequences that are repeated over and over again without interruption along a segment of a chromosome. For example, the normal allele of the fragile X gene contains from about 5 to about 50 tandem repeats of the trinucleotide CGG.

Section II: Computer Terms

bit: the smallest unit of information that a computer can hold.

byte: a unit of information consisting of a fixed number of bits. One byte usually consists of a series of eight bits.

click and drag: a technique used to choose items from a menu or to move an object. To click and drag, position the pointer on an object or menu item, press the mouse button, and hold it down while moving the mouse to a new position.

close box: the small box on the left end of the title bar of an active window. Clicking on a close box closes the window.

database: any collection of information that is organized in a specific way.

default: on a computer, any value that appears automatically without active user choice.

desktop: the working environment on a computer. The desktop includes the menu bar and the background area on the screen in which a user works with icons and windows.

double-click: to place the pointer on an object (such as an icon) and then to press and release the mouse button two times in quick succession without moving the pointer.

electronic database: a computer, the data, and the software required to store, manipulate, and display those data.

gigabyte: 1.0737 x 10^9 bytes.

icon: a small pictorial representation of a file, desk, menu, option, or other object or feature.

informatics: the study of the storage and management of data.

Internet: a worldwide umbrella group for more than 7,500 electronic mail networks that have agreed on standard policies and procedures for physically transmitting electronic mail.

kilobyte: 1.024 bytes.

megabyte: 1.048 x 10^6 bytes.

menu bar: a strip across the top of a computer screen that contains the names of the menus available to the user.

online database: remote electronic database that can be accessed using computers linked by telecommunications.

pointer: in a computer, an arrow or other symbol on the screen that moves as a user moves the mouse; in a database, anything that refers the user to the item in question (for example, in a card catalog, the author card is a pointer to the actual book).

pop-up box: a box that appears after making a choice from a menu or clicking on an item. A pop-up box may give additional information or require additional input.

pull-down menu: a menu (usually in the menu bar) whose name or icon always is shown. Users can "pull down" the menu by placing the pointer on the name or icon, pressing and holding the mouse button, and moving the mouse down to expose the menu.

RAM: Random Access Memory. The part of a computer's memory available for programs and documents.

registry database: a database that contains various types of information about individual people. Registry databases contain information identified as "belonging to" or "describing" specific individuals.

research database: a database that stores the results of research. HGP-related research databases store information about the generalized human genome. Genetic information stored in a research database cannot be identified as "belonging to" or "describing" specific individuals.

scroll-bar arrow: arrow used to move a document or directory in its window so that a different part is visible.

terabyte: 1.0995 x 10^{12} bytes.

title bar: the bar at the top of a window that shows the name of the window. When the window is active, the title bar is highlighted with horizontal lines.

window: a rectangular area on a computer screen that displays information. Users create and view documents through windows.

SECTION III: GENETIC DISORDERS

cystic fibrosis (CF): the most common hereditary disorder in the Caucasian population. In the United States, the frequency of individuals affected with CF is about 1/2,000; heterozygotes make up approximately 5 percent of the population. The CF gene encodes a protein (containing 1,480 amino acids) that is thought to control the transport of chloride ions through cell membranes.

Duchenne muscular dystrophy (DMD): a usually fatal disorder involving a progressive weakening and wasting of the muscles. DMD is inherited as an X-linked recessive trait and occurs in about 1/3,600 live male births. The DMD gene, consisting of approximately 2 million base pairs and 60 exons, is about 10 times larger than any other known gene. Most of the reported alterations of the DMD gene are deletions or duplications of different portions of the genetic material.

familial hypertrophic cardiomyopathy: a common form of inherited cardiomyopathy (a cardiomyopathy is a disease that primarily affects the myocardium of the heart) and an important cause of sudden death in the young. The condition is inherited as an autosomal dominant trait. Symptoms include breathlessness and angina, and there is a considerable risk of sudden death due to arrhythmias. The molecular basis of familial hypertrophic cardiomyopathy in some patients is a mutation in one of the cardiac myosin heavy-chain genes.

fragile X syndrome: a set of symptoms (including mental retardation and various physical characteristics) that is associated with a specific fragile site on the X chromosome. Fragile X syndrome is inherited as an X-linked dominant trait with reduced penetrance (that is, not all individuals carrying the gene necessarily show signs of the condition). Scientists studying the gene have traced the condition to an increase in the number of CGG repeats that occur in one portion of the gene. Fragile X syndrome occurs in all ethnic groups with equal frequency and is the most common inherited condition associated with mental retardation.

hereditary juvenile glaucoma (HJG): an autosomal dominant condition that has its onset between the ages of 3 and 20. Glaucoma is characterized by increased pressure from aqueous humor within the eye, leading ultimately to blindness if untreated. HJG varies in expression from family to family.

Huntington disease (HD): a heritable condition involving progressive degeneration of the central nervous system. The condition usually manifests itself in a person's 30s or 40s as small involuntary movements (chorea) that gradually worsen to involve all parts of the body. Problems with memory and thinking also may occur. HD is inherited as an autosomal dominant trait. Scientists studying the gene have traced the condition to an increase in the number of CAG repeats that occur in one portion of the gene. The number of repeats relates to the severity of the disease as well as to the age at which the first symptoms appear (the greater the number of repeats, the more severe the symptoms and the earlier the onset).

Kennedy disease: a condition resulting from the enlargement of a tandem CAG repeat within the first exon of the androgen receptor gene (located on the X chromosome). Symptoms include muscle weakness and wasting, and, in some patients, sensory abnormalities. The condition usually is not fatal.

myotonic dystrophy: a usually progressive, inherited disorder characterized by weakening of the muscles and other medical problems, which may include cataracts, diabetes, irregular heart beat, and learning disabilities. The mutated allele of the myotonic dystrophy gene shows an increase in the number of times the trinucleotide CTG is repeated.

phenylketonuria (PKU): a condition associated with a defect in phenylalanine metabolism. The resulting accumulation of excess phenylalanine can result in abnormal neurological development and accompanying mental retardation if a low protein diet and a formula supplementation are not implemented.

sickle cell disease (SCD): a generally fatal form of hemolytic anemia that occurs in individuals who are homozygous for the abnormal allele H^S. The red blood cells of such individuals contain an abnormal hemoglobin that causes them to undergo a reversible change in shape when the oxygen level of the plasma falls slightly. Such cells have a greatly reduced life span, because they tend to clump together and are rapidly destroyed. The incidence of sickle cell disease in African-Americans in the United States is approximately 1/500.

spinocerebellar ataxia: a condition marked by an increasing inability to coordinate muscle action or to control involuntary movement. The symptoms appear to be related to degeneration of the cerebellum and other parts of the nervous system. Clinical expression and age of onset are highly variable. One form of the condition has been traced to expansion of the trinucleotide CAG.

References

Andrews, L.B., et al. (1994). *Assessing genetic risks: Implications for health and social policy.* Washington, D.C.: National Academy Press.

Annas, G. (1993, November). Privacy rules for DNA databanks. *Journal of the American Medical Association* 270(19):2346-2350.

Beauchamp, T.L. and Childress, I.F. (1989). *Principles of biomedical ethics.* 3rd edition. New York: Oxford University Press.

Bennet, R.L., Steinhaus, K.A., Uhrich, S.B., et al. (1995). Recommendations for standard human pedigree nomenclature. *American Journal of Human Genetics* 56:745-752.

Billings, P.R., Kohn, M.A., deCuervas, M., Beckwith, J., Alper, J.S., II, and Natowicz, M.R. (1992, April). Discrimination as a consequence of genetic testing. *American Journal of Human Genetics* 50:478-479.

Bishop, J.E. and Waldholz, M. (1990). *Genome.* New York: Simon and Schuster.

Brown, R.S. and Marshall, K. (eds.). (1993). *Advances in genetic information: A guide for state policy makers.* 2nd edition. Kentucky: The Council of State Governments.

Cuticchia, A.J., Chipperfield, M.A., Porter, C.J., Kerns, W., and Pierson, D.L. (1993, October 1). Managing all those bytes: The Human Genome Project. *Science* 262(5130):47-48.

Chaum, D. (1992, August). Achieving electronic privacy. *Scientific American* 267(2):96-101.

Collins, F. and Galas, D. (1993). A new five-year plan for the U.S. Human Genome Project. *Science* 262(5130):43-46.

Cooper, N.G. (ed.). (1992). *Los Alamos Science #20.* Los Alamos, NM: Los Alamos National Laboratory.

Erickson, D. (1992, April). Hacking the genome. *Scientific American* 266(4):128-137.

Feinberg, J. (1984). *Harm to others.* New York: Oxford University Press.

Fraser, C.M., Gocayne, J.D., White, O., et al. (1995). The minimal gene complement of *Mycoplasma genitalium. Science* 270:397-403.

Holtzman, N.A. (1989). *Proceed with caution: Predicting genetic risks in the recombinant DNA era.* Baltimore, MD: The Johns Hopkins University Press.

Kevles, D.J. and Hood, L. (eds.). (1992). *The code of codes: Scientific and social issues in the Human Genome Project.* Cambridge: Harvard University Press.

Lee, T.F. (1991). *The Human Genome Project: Cracking the genetic code of life.* New York: Plenum Press.

Nelkin, D. and Lindee, M.S. (1995). *The DNA mystique: The gene as a cultural icon.* New York: W.H. Freeman & Company.

Privacy Protection Study Commission. (1977). *Personal privacy in an information society.* Washington D.C.: Privacy Protection Study.

Reich, W.T. (ed.). (1987). *Encyclopedia of bioethics*. New York: The Free Press.

Strohman, R.C. (1993). Ancient genomes, wise bodies, unhealthy people: Limits of a genetic paradigm in biology and medicine. *Perspectives in Biology and Medicine* 37(1): 112-145.

Task Force on Genetic Testing. (1996). *Interim Principles.* NIH-DOE Working Group on the Ethical, Legal, and Social Implications of Human Genome Research.

U.S. Congress, Office of Technology Assessment. (1990). *Genetic monitoring and screening in the workplace.* OTA-BA-455. Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Health and Human Services and U.S. Department of Energy. (1990, April). *Understanding our genetic inheritance. The U.S. Human Genome Project: The first five years, FY 1991-1995.* Springfield, VA: National Technical Information Service.

U.S. National Institutes of Health and U.S. Department of Energy Working Group on Ethical, Legal, and Social Implications of Human Genome Research. *Genetic information and health insurance.* Report of the Task Force on Genetic Information and Insurance. NIH Publication No. 94=3-3686, May 10, 1993.

The White House Domestic Policy Council. (1993, draft version). *The president's security plan: The Clinton blueprint*. New York: TimesBooks.

Additional Resources

You can obtain additional information on the genetic disorders listed below by contacting the support organizations indicated. In most cases, this information will be mailed to you free of charge.

Cystic Fibrosis Foundation
6931 Arlington Road #200
Bethesda, Maryland 20814
(301) 951-4422

Duchenne Muscular Dystrophy Association
3300 East Sunrise Drive
Tucson, Arizona 85718
(602) 529-2000
e-mail: 74431.2513@compuserve.com

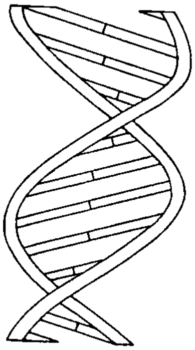
National Fragile X Foundation
1441 York Street, Suite 303
Denver, Colorado 80206
(303) 333-6155
(800) 688-8765
e-mail: natfragx@ix.netcom.com

Foundation for Glaucoma Research
490 Post Street, Suite 830
San Francisco, California 94102
(415) 986-3162
home page: www.glaucoma.org

Huntington Disease Society of America
140 West 22nd Street, 6th Floor
New York, New York 10011-2420
(212) 242-1968
e-mail: h.d.s.a.ttisms.com

PKU Parents
Care of Dale Hillard
8 Myrtle Lane
San Anselmo, California 94960
(415) 457-4632

Sickle Cell Disease Association of America
200 Corporate Pointe, Suite 495
Culver City, California 90230-7633
(310) 216-6363
(800) 421-8453



Introductory Activity The HGP and Electronic Databases

Focus

In this activity, students compare the results of a manual search of DNA sequence data with the results they obtain using the computer. The activity introduces students to the Human Genome Project (HGP) and illustrates the usefulness of electronic storage and search techniques in handling large amounts of genomic data.

HGP Context

Research scientists predict that the HGP will generate more data than any other single research effort in the history of biology. Simply to record the complete sequence of the haploid human genome will require the equivalent of 200 telephone books of 1,000 pages each. Variation in the sequences among humans as well as information about the genomes of other research organisms will increase further the amount of data that must be stored.

The data collection, storage, and distribution requirements of the HGP are such that government agencies have allocated significant funding to the design and development of electronic databases to store genomic information. Electronic storage offers scientists many advantages over print-based storage systems. For example, electronic databases require less space to store the data, allow the data to be kept consistent and accurate more easily, and permit rapid searches that automatically identify all relevant records according to multiple criteria. These advantages are vital to scientists who must deal with a vast body of complex information that continues to grow daily.

Major Concepts

□ The HGP will generate map and sequence data at a rate and volume that are of different orders of magnitude from that which biologists have experienced previously. Scientists and information specialists must collect, store, analyze, and make this information accessible to other researchers if it is to be fully useful.

□ Electronic genomic databases are powerful resources that can help us answer biological questions, especially questions about gene structure and function, the mechanisms of genetic disease, and similarities and differences among individual organisms and species.

Student Objectives

As students complete this activity, they should

□ understand that the volume and complexity of genomic data are such that scientists must use electronic databases to organize, store, and search them effectively.

Science Process Skills

□ Observing
□ Comparing
□ Communicating (orally and in writing)
□ Using the computer as a research tool

Materials for a Class of 30

□ 30 copies of student text, p. S-1. Because students will complete their written work on their own paper or on the worksheets provided, other students can reuse these pages.
□ 30 copies each of Worksheet 1, *Introduction to the Human Genome Project*; Worksheet 2, *Finding a Sequence*; and Worksheet 3, *Genomic Databases*. Assign Worksheet 3 as homework in preparation for Activity 1.
□ Overhead transparency of BLM T-1, *Target Sequences*.

Advance Preparation

□ Install the software onto computers for students to use. Set the software on each computer to *The HGP and Electronic Databases*. You will find instructions for how to do this on pp. 55–56. The password is <Watson>.
□.Make the necessary student copies and overhead transparencies.
□ Two days before: Ask your students to read the background information for the introductory activity (Worksheet 1, *Introduction to the Human Genome Project*) and answer the study questions provided.

Estimated Time

30 minutes

Introduction

As the volume and the complexity of information about the human genome have increased, the technologies required to store, communicate, and analyze this information also have changed. At the First International Human Gene Mapping Workshop in 1973, 75 participants summarized the human map on one printed page, which contained the names of 25 genes. By early 1996, records on more than 5,900 human genes were stored in the Genome Database, which is a national, computer-based repository for human gene mapping information, and more than 460 million bases of DNA sequence data were stored in GenBank, another national genomic database.

This activity introduces students to some of the problems associated with storing, retrieving, and analyzing the enormous volume of genetic data that the HGP is generating. The activity also illustrates how scientists can use electronic databases to help them retrieve sequence information quickly and accurately.

Related Information in the Teacher Narrative

You will find additional background information in the following pages of the teacher narrative:
□ objectives of the HGP (pp. 1–3)
□ rate of increase in mapping and sequencing data (p. 17)
□ techniques used to identify genes (pp. 9–14)
□ storing and using sequence data (pp. 20–21)

Suggestions from the Field Test

□ Use an overhead transparency made from Table 2 in the teacher narrative (p. 17) to illustrate the rapid increase in the volume of information stored in three major research databases associated with the HGP.
□ Have the students complete each of the three manual searches using a different colored pencil so that they can compare the results of each search more easily.

76

Annotated Student Activity

This activity assumes that students have read the background material we provide (Worksheet 1, *Introduction to the Human Genome Project*). The activity introduces students to some of the difficulties associated with print-based storage and retrieval of sequence data and illustrates the computer's ability to handle such data quickly and accurately.

Before you begin, divide the class into work groups around the available computers. (Students should remain in the same work groups for all of the activities in the module. You will find suggestions for how to organize your groups on pp. 49–51.)

Imagine that you are a member of a team of scientists that is searching for a particular gene. You have just sequenced a short section of DNA that you think might be a piece of the gene, and now you need to search the published sequence data to see whether anyone else already has reported the same sequence.

PROCEDURE

1. Follow your teacher's instructions for completing a manual search (a search by hand) of the sequence data that you have been given.

Distribute one copy of Worksheet 2, *Finding a Sequence*, to each student. Explain that the As, Cs, Ts, and Gs on the sheet represent the base sequence of a short section of DNA. Ask the students to perform manually the searches described below by using only the worksheets with the printed sequence and their pencils. Project the target sequence for each of these searches using an overhead transparency (BLM T-1); use a piece of paper to cover the sequences and reveal them one at a time as appropriate.

a. To give the students a sense of what they are to be doing, ask them to find and circle all of the AGG sequences on their sheets (Target Sequence 1), reading from left to right. (The AGG sequence represents the first three bases in the new DNA that they just identified; the bases on the printed page represent the sequence data already published in the scientific literature.)

Explain to the students that although the printed page allows only a limited number of letters to be displayed on one line, an actual DNA sequence would be read continuously from one end to the other. Consequently, they should assume that each line of sequence continues to the next without a break, and they should circle even those AGG sequences that span the end of one line and the beginning of the next. Point out, for example, that the bases A and G at the end of the first line of the sequence, together with the base G at the beginning of the second line, constitute a valid occurrence of the sequence AGG and should, therefore, be circled.

The numbers along the left margin of the sequence are base counts. The sequence shown contains a total of 2,508 bases and represents a modified section of the actual sequence of the human APRT gene. This gene encodes the enzyme adenine phosphoribosyltransferase, which participates in salvage reactions that convert free nitrogen bases produced by the breakdown of DNA or RNA to the corresponding nucleotides. The complete sequence for the APRT gene contains approximately 3,000 bases.

Give your students one minute in which to accomplish this search. When the minute is over, ask them to count the number of AGG sequences they found during that time and to record the number on their own paper. They should identify this number by the description "Manual Search, Target Sequence 1." (The answer sheet for Worksheet 2 is shown on p. 73. Your students likely will not have identified all of the AGG sequences during their one-minute searches.)

b. Next, ask students to find all of the AGGGGA sequences that appear on the worksheet (these are the first six bases of the new DNA that they just identified). Remind them to read from left to right across each line and also to read continuously from one line to the next. Give them another minute in which to accomplish this second search. After the minute is up, ask them to count and record the number of AGGGGA sequences they found (Manual Search, Target Sequence 2).

c. Finally, ask them to find the full sequence of their new DNA: AGGGGAGTCAGGGGCTCTGCATGAGGAGGG. Give them one minute to accomplish this. Again, ask them to record the number of full sequences they found (Manual Search, Target Sequence 3).

d. Point out to students that the printed sequence on the worksheet contains only about 2,500 bases and remind them that in early 1996, GenBank contained more than 460 million bases of reported sequence. The average human gene is about 20,000 bases, although some are much longer. The gene for cystic fibrosis, for example, is about 250,000 bases and the gene for muscular dystrophy is about two million bases. The complete human sequence contains more than three billion bases.

Point out as well that the sequences that they were searching for also were very short. A scientist conducting a more typical search of the sequence literature would attempt to match several hundred or several thousand bases of new sequence against all of the data stored in a sequence database.

2. Follow your teacher's instructions to bring up the sequence data stored in your computer. Complete the steps listed in the box below to repeat the searches you just conducted manually, this time having the computer do the searching.

The computers already should be loaded with the required software and the activity set to *The HGP and Electronic Databases*. You will find instructions for completing this on pp. 55–56.

To Search the Sequence Data for a Target Sequence

- Click and pull on the box labeled *TARGET SEQUENCE* to highlight the sequence of interest. (**Windows Users:** Click on the arrow.)
- Release the mouse button.
- Click on *BEGIN SEARCH.*

NOTE: Click on the scroll bar arrow at the right of the screen to see the whole sequence.

a. Search the sequence data for the first target sequence (AGG). What data does the computer retrieve for you?

The computer shows all of the AGGs in bold and underlined.

b. Click on *COUNT* to determine the total number of AGG sequences that the computer found. Record this number and compare it with the number that you found during your manual search.

The computer locates 78 AGG sequences. This number should be much higher than the numbers the students found during their manual searches.

c. Now search for the slightly longer sequence, AGGGGA. Again, click on *COUNT* to determine the number of matching sequences that the computer found. Record this number and compare it with the number of matches that you found during your manual search.

There are three AGGGGAs in this sequence. It is unlikely that any students will have found all three occurrences during the manual search.

d. Finally, select the full sequence of interest: AGGGGAGTCAGGGGCTCTGCATGAGGAGGG. How many times does this sequence occur in your data? How long did it take the computer to locate the sequence?

The sequence occurs once in these data. The computer located the sequence within the first few seconds of search time. It is unlikely that the students found this sequence in their manual searches even though they were given a full minute during which to look.

Questions for Discussion

1. How does the difficulty of the search change as your target sequence gets longer?

The search becomes more difficult as the target sequence gets longer. It also becomes more difficult as the amount of sequence data that must be *searched* gets larger.

2. What do the results of the two different types of searches suggest about the advantages of computer-based storage and retrieval of sequence data?

Computer-based retrieval of sequence data is more rapid than manual search techniques. Computer-based searches of sequence data also are more accurate than manual searches. For example, the first 29 bases of the longest target sequence also occur starting at base 886 (position 27 of the line beginning with base 859), but the last base (the base at position 56 of the same line) does not match. The computer correctly did not bold this incomplete sequence. Computer-based storage also requires less space than print storage. These advantages explain why the HGP is so heavily dependent upon computers, not only for data storage and dissemination, but also for data analysis and manipulation.

Homework Assignment

Read the background information for Activity 1 (Worksheet 3, *Genomic Databases*) and answer the study question provided.

Annotations to Worksheet 1
Introduction to the Human Genome Project

Your students may wonder why the estimate of the total number of human genes spans such a wide range (50,000-80,000 genes). Explain that such estimates are based on certain assumptions scientists make about the total number of bases that make up the human genome (approximately three billion), the percentage of the genome that codes for specific gene products or functions (about 5 percent), and the average length of a human gene (about 20,000 bases). These assumptions lead to an estimate of approximately 300,000 human genes, an estimate that scientists narrow down to 50,000-80,000 based on their knowledge of the genomes of a variety of other organisms.

1. When biologists talk about the human genome, they often use terms such as *gene*, *chromosome*, and *base* to refer to different levels of organization of the genetic material. Using each of these terms correctly, describe what scientists must accomplish in order to map and sequence the human genome. Be sure that your answer illustrates clearly the relationships that exist among each of these levels of organization.

To map the human genome, scientists will have to determine the precise location of each of the estimated 50,000 to 80,000 human *genes* within the 25 different types of *chromosomes* that make up the genome (22 autosomes, plus the X, Y, and mitochondrial chromosomes). To sequence the human genome, scientists will have to determine the order of the estimated three billion *bases* (As, Ts, Cs, and Gs) along the DNA that makes up each of these 25 different types of chromosomes. This means determining the order of the bases that make up each gene, as well as the order of the bases that make up the DNA that lies between genes.

2. You probably have learned that most human cells contain 46 chromosomes (notable exceptions include eggs and sperm). You also probably have learned that different forms of the same gene—for example, different *alleles* of the gene for cystic fibrosis —contain different sequences of bases.

a. Explain how it is, then, that scientists studying the human genome will map only 25 different types of chromosomes.

Because the order of genes along the two members of a pair of homologous chromosomes typically is the same, scientists involved in mapping the genome only have to study one of each type of human autosome (that is, 22 of the 44 autosomes in a diploid human cell), plus the X, Y, and mitochondrial chromosomes.

b. What do you think scientists studying the human genome mean when they talk about sequencing "the" human genome? Does "the" human sequence actually exist?

Because people differ in the precise order of the bases in their genomes, there really is no *single* "human" sequence. As scientists determine the base sequence of more and more human DNA, they document all of the sequence differences that they discover in the DNA taken from different individuals. This means that scientists involved in the HGP really are not sequencing "the" human genome, as such. Instead, they are documenting the scope of human variability *at the molecular level.* The complete human sequence will be a composite of sequences derived from many individuals.

3. Although the focus of the HGP is on mapping and sequencing the human genome, scientists involved in the effort also are studying the genomes of several other organisms.

a. Table 1 lists the approximate sizes of the genomes of several of the HGP organisms. If you assume an average sequencing rate of 10 million bases each year, how long would it take to sequence each of these genomes?

Table 1 Approximate sizes of the genomes of several HGP organisms.

Type of Organism	Estimated Numbers of Base Pairs (millions of pairs)
bacterium	4.7
yeast	15
fruit fly	80
human	3,000

The purpose of this question is to help students see that sequencing these genomes is a large task and to give students some sense of the relative challenges involved in sequencing genomes of different sizes. Assuming an average sequencing rate of 10 million bases each year, genomes of the sizes indicated could be sequenced in .47, 1.5, 8, and 300 years respectively. At the beginning of 1996, the average sequencing rate worldwide was estimated to be between 10 and 100 million bases per year. Even if we assume the fastest rate, it would take .047, .15, .8, and 30 years respectively to sequence genomes of the sizes indicated.

b. The background information we provide identifies some of the benefits that scientists are realizing by mapping and sequencing these nonhuman genomes. What do these benefits tell us about the organization and function of these nonhuman genomes in comparison with the organization and function of the human genome? What do they tell us about evolution?

The general organization and function of nonhuman genomes are similar to those of the human genome. DNA is the universal information molecule in the living world (save RNA viruses), and the mechanisms for replication, transcription, and translation are universal as well. This universality of structure and function in the genetic material provides compelling evidence that all life is related through descent with modification. Indeed, the evolutionary history of any species is written largely in its genes.

ANNOTATIONS TO WORKSHEET 2
Finding a Sequence

1 CTACGGTGACAGCTGCC**AGG**ATCCTAAA**AGG**GCAGAAGA**AGG**ACAAACTGGGGCCTGAGACCTT**AG**

67 **G**GGCCATGGACCGCTTCCCGTACGTGGCTCTGTCCA**AGG**TAAGTGCTGGGCTACCTTAGAGTCCTC

133 CAAGCAGAGA***AGGGGA***ATCCTGGCTATGGAGTGTGGT**AGGAGG**G**AGG**GACCCTAAACAGCTGGGGC

199 TCCAATA**AGG**AGCTGG**AGG**CAGTTGGAATCCCAG**AGG**ACAGAGATC**AGG**GTCTTGTTTGTCTGCCC

265 CAGAGAAGAGCTCAGAGTGTCTCTGTCCCCAGACATACAGTGTAGACAAGCATGTGCCAGACAGTG

331 GAGCCACAGCCACGGCCTACCTGTGCGGGGTCA**AGG**GCAACTTCCAGACCATTGGCTTGAGTGCAG

397 CCGCCCGCTTTAACCAGTGCAACACGACACGCGGCAACG**AGG**TCATCTCCGTGGTGAATCGGGCCA

463 AGAAAGC**AGG**TGGAGCTGGGGCCCGGCTGTGGGGTC**AGG**GCCAGTGACAGACCTCTATCGCATATC

529 CTGACCTCTATCACCCTC**AGG**AAAGTCAGTGGGAGTGGTAACCACCACACGGGTGCAGCATGCCTC

595 GCCAGCCGGCACCTACGCCCACACGGTGAACCGCAACTGGTACTCGGATGCCGACGTGCCTGCCTC

661 GGCCCGCC**AGGAGG**GGTGCC**AGG**ACATCGCCACGCAGCTCATCTCCAACATGGACATTGATGTGCG

727 ACCCCCGGGCCA**AGG**GTGGGCTGGCAGAGAGTAGC**AGG**G**AGG**GGCACCAGCTCAGACC**AGG**CA

793 ACCAAAAGCCTTATCTGGGCCAGC**AGG**GTCTGGA**AGG**TGGGGTTGGGGGCGTAGA**AGG**CGCACC**AG**

859 **G**CTGGGCCATTCCCACAGCCTTGGG***AGGGGA***GTC**AGG**GGCTCTGCATG**AGGAGG**TGACACGGGGC

925 CTAGCCATGGCCCAAAGTCCACCTGCCCCATCCTCTGTTCCC**AGG**TGATCCT**AGG**TGG**AGG**CCGAA

991 AGTACATGTTTCCCATGGGGACCCCAGACCCTGAGTACCCAGATGACTACAGCCA**AGG**TGGGACC**A**

1057 **GG**CTGGACGGGAAGAATCTGGTGC**AGG**AATGGCTGGCGAAGCACC**AGG**TGATGGGGGCTGGTGGGT

1123 GTGCTGGGCACAGC**AGG**GGG**AGG**GCAG**AGG**TGTGGGCTCGGGGCTGTGGGCTG**AGG**CCTGGCTCT

1189 CTCCCTCCCCGC**AGG**GTGCCCGGTACGTGTGGAACCGCACTGAGCTCCTGC**AGG**CTTCCCTGGACC

1255 CGTCTGTGACCCATCTCATGGGTAATGACCCCCTTCCTGCCCTGGCATCCTCAGATGGCCTCAGAT

1321 GGCACTTCTGAGCCTGTGTGCACATCCGCCAGCACCCTCCCACCCCCAGCCTGCCAGTCACCAC**AG**

1387 **G**ACCCCTTGTCCCAC**AGG**TCTCTTTGAGCCTGGAGACATGAAATACGAGATCCACCGAGACTCCAC

1453 ACTGGACCCCTCCCTGATGGAGATGACAG**AGG**CTGCCCTGCTCCTGCTGAGC**AGG**AACCCCCGCGG

1519 CTTCTTCCTCTTCGTGG**AGG**GTGCGTGGTGGCCCTGGGAGTGGGGGGTTGGGGGTTGGAGC**AGG**GC

1585 **AGG**CTCAGCATCTCCCCCCTCTGGCCTTCCTGC**AGG**TGGTCGCATCGACCATGGTCATCATGAAAG

1651 C**AGG**GCTTACCGGGCACTGACTGAGACGATCATGTTCGACGACGCCATTGAG**AGG**GCGGGCCAGCT

1717 CACCAGC**GAGGAGG**ACACGCTGAGCCTCGTCACTGCCGACCACTCCCACGTCTTCTCCTTCGG**AGG**

1783 CTACCCCCTGCG**AGG**GAGCTCCATCTTCGGT**AGG**CCTGGGGATGAGTGGC**AGG**TGCTGCTGCAGCA

1849 ATTAAGTGGGTGAAATCTGAGCCTCAGTCTCCTCCTCTGTCAAGTGGGAGTAATGCTGGCACCAGC

1915 CTAAT**AGG**GTCCTCTGCGGACTAAGCCCCTGACC**AGG**CAAAACGTGCGGTGCCTAGCACGTGGGAG

1981 ACACTCCACAGCTGTGTTCAGCTCAACCAC**AGG**GACCCCTCTCTC***AGGGGA*GTCAGGGGCTCTGCA**

2047 **TGAGGAGGG**C**AGG**A**AGG**CCTACACGGTCCTCCTATACGGAAACGGTCC**AGG**CTATGTGCTCA**AGG**A

2113 CGGCGCCCGGCCGGATGTTACGGAGAGCGAGAGCGGTGAGTGCCGTGGGGTGGCCTG**AGG**GGGACC

2179 **AGG**GTGCCA**AGG**ATGGGGGGCTGGCGGGA**AGG**GGTCACCTCTTGTCTGCCTGGAACTGAAACTTCC

2245 TACTGAAACTGAACCCTCCAACC**AGG**GAGCCCCGAGTATCGGCAGCAGTCAGCAGTGCCCCTGGAC

2311 GGAGAGACCCACGC**AGG**CG**AGG**ACGTGGCGGTGTTCGCGCGCGGCCCGC**AGG**CGCACCTGGTTCAC

2377 GGCGTGC**AGG**AGCAGACCTTCATAGCGCACGTCATGGCCTTCGCCGCCTGCCTGGAGCCCTACACC

2443 GCCTGCGACCTGGCGCCCGGCACTTCTGAGCCTGTGTGCACATCCGCCAGCACCCTCCCACCCCCA

Key: AGGs are shown in bold and underlined. *AGGGGA*s are shown in bold and are underlined and italicized. AGGGGAGTCAGGGGCTCTGCATGAGGAGGG is shown in bold and underlined. Note that students might find the sequence *AGGGGA*GTCAGGGGCTCTGCATGAGGAGGT in the line starting with base 859. This sequence matches the target sequence only for the first 29 bases; the 30th base in the target sequence is a G rather than a T.

ANNOTATIONS TO WORKSHEET 3
Genomic Databases

Use the information provided in the reading to develop your own definitions of the following terms:

Because students are not given formal definitions of these terms, their answers will vary. Answers should include the elements listed below.

a. database

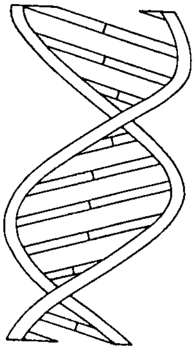
Answers should include the idea that scientists use databases to collect and organize information to allow easy retrieval.

b. research database

Answers should include the ideas that information in a research database *is not* identified as being about any one specific person and that this information helps scientists answer questions of general biological interest.

c. registry database

Answers should include the idea that information in a registry database *is* identified as belonging to and describing specific individuals and therefore can be used to answer questions about those individuals.



Activity 1
Genetic Registries

Focus

This activity introduces students to the model of a registry database that accompanies this module. Students identify with specific individuals in a set of three fictitious families, use simple search strategies to locate information about these individuals in the database, and share that information to create pedigrees for each of the extended families. In preparation for Activity 4, *Who Should Control Information about My Genes?*, students also decide whether to authorize having their fictitious person tested for his or her genotype for each of three additional genes.

HGP Context

The biology of the Human Genome Project (HGP) and general issues related to database construction and use are important interrelated concepts. Because the final product of the HGP will be a complete physical map of the human genome and because these data will be stored and accessed electronically, it is critical for all of us to consider questions related to the development, use, and regulation of genomic databases.

With respect to the HGP, we can define two major types of databases. *Research databases* allow investigators to store and manipulate information that is not identified as relating to specific individuals. National research databases such as Genome Database (GDB), GenBank, and Online *Mendelian Inheritance in Man* (OMIM) contain information about the names, functions, map positions, sequences, and clinical significance of identified human genes. Hundreds of research laboratories around the world contribute this information. Because of the persistent efforts of those involved in collecting and verifying these data, the information stored in these databases approximates the current state of our knowledge and understanding of the generalized human genome at any point in time. Activities 2-4 require students to interact repeatedly with a model of a research database that we call the National Genome Database (NGD). The NGD combines the types of information that are found in a variety of existing research databases.

Registry databases, the second type of database used in the HGP, contain information on specific individuals. Because registries may include sensitive personal data such as health histories, genetic profiles, and the results of various medical or psychological tests, the individuals responsible for creating and maintaining these databases usually limit access to this information to authorized users. Activities 1-3 require students to retrieve data from a model genetic registry that we call the Local Genome Database (LGD). The LGD combines elements of information typically stored by hospitals, insurance companies, employers, and schools. In Activities 1-3, the students access these personal data freely. In Activities 4 and 5, the students examine some of the ethical issues associated with access to such data and consider some of the public-policy questions that arise from these issues.

Major Concepts

□ "The" human genome does not exist. Although we can describe a generalized human genome, occasional differences in map position, as well as variations in particular sequences (allelic differences), allow DNA from each of us to be identified as unique.

□ The scientific community stores genomic data in two types of databases. *Research databases* organize and store aggregated information about the generalized human genome (for example, the names of genes, map positions, clinical information, and lists of DNA and protein sequences). *Registry databases*, on the other hand, store personal genomic data that can be used to describe or to identify individuals by their specific genetic profiles.

□ The content and the structure of a database limits to some extent the types of questions one can ask, but the usefulness of a database also can be limited by the types of questions the researcher asks.

Student Objectives

As students complete this activity, they should

□ understand that registry databases store information that is identified as belonging to particular individuals,

□ search a registry database for personal genomic data relating to a fictitious individual, and

□ construct a pedigree for that individual's family on the basis of relationships reported in the registry.

Science Process Skills

□ Gathering data

□ Synthesizing information and knowledge

□ Communicating (orally and in writing)

□ Using the computer as a research tool

Materials for a Class of 30

□ 30 sets of the student text, pp. S-7–S-9. Because students will complete their written work on their own paper or on the worksheets provided, other students can reuse these pages.

□ 30 copies each of Worksheet 4, *Collecting Family Data*; Worksheet 5, *Constructing a Pedigree*; Worksheet 6, *Interpreting Data in the LGD*; and Worksheet 7, *Kate and Ryan*. Assign Worksheet 7 as homework in preparation for Activity 2.

□ 1 copy of BLM T-2, *Assigning Fictitious Identities* for each class that will be completing Activity 1.

□ 90 cotton-tipped swabs and 30 small envelopes.

Advance Preparation

□ Homework: Ask your students to read the background material for Activity 1 (Worksheet 3, *Genomic Databases*) and answer the study question provided. (Worksheet 3 is located on pp. S-5–S-6; the answers are located on p. 74.)

□ Make the necessary student copies.

□ Make 1 copy of BLM T-2, *Assigning Fictitious Identities*, for each class that will be completing Activity 1. Follow the instructions provided on the worksheet to assign a fictitious identity to each student in each class.

□ Set the software on each computer to *Genetic Registries*. You will find instructions for how to do this on p. 56. The password is <Lyon>. You may wish to have your students activate the appropriate data; we provide a blackline master listing the steps required (BLM T-3).

□ The last step in Activity 1 (Step 5) sets the stage for Activity 4, *Who Should Control Information about My Genes?* You may wish to read Activity 4 before you ask your students to complete this step.

Estimated Time

Two 45-minute class periods

Introduction

Activity 1 introduces students to the Local Genome Database (LGD), the model registry database developed for this module. Throughout the activity, students will move back and forth among various files in the LGD to gather and analyze specific health and genetic information about "themselves" and their fictitious families.

You may find that some of your students are confused about the difference between the consensus sequences that are stored in a research database (for example, the NGD that students will encounter in Activity 2) and the personal sequences that are stored in a registry database like the LGD. Explain to your students that investigators are not mapping and sequencing any *one person's* genome. Instead, they are studying human DNA prepared from a variety of sources. Because the chromosomal locations of specific genes usually do not vary from one person to the next, map positions determined from one source are applicable to all humans. However, because the actual base sequences of specific genes often vary among humans (that is, because genes show allelic variation), the base sequence of a gene derived from one person may be different from the base sequence of the same gene derived from another. These individual differences are illustrated by the base sequence differences among individuals in the LGD.

Emphasize the importance of these polymorphisms to your students by pointing out that these sequence differences, in part, distinguish each of us from the other and by reminding them that such polymorphisms also allow us to study the transmission of genetic information. Explain that a research database stores information about all of the different allelic variations that scientists find and that these data are anonymous (that is, the genetic information in a research database *is not* linked to specific individuals). In contrast, a registry database contains only those allelic variations found among the individuals listed in the registry, and these data are not anonymous (that is, the genetic information in a registry *is* linked to specific individuals).

Unlike most registry databases, the LGD is not protected by a password. Instead, the LGD is structured so that students can access freely the records of all of the members of each fictitious family. The pedagogical goals here are 1) to allow the students to explore the full range of the personal data that are contained within the LGD, and 2) to allow the students to discover for themselves some of the reasons that access to information in registry databases typically is restricted to authorized individuals.

Privacy and confidentiality, then, are *not* central issues in Activity 1, and we recommend that you do not raise them. Instead, encourage your students to search widely through the information in the LGD and invite them to discuss that information freely among themselves.

In contrast, the exercises in Activities 2 and 3 are designed to alert students to a number of increasingly compelling reasons to be concerned about the accuracy and the privacy of genomic information that is stored in genetic registries. In Activity 2, students see that databases can contain confusing or erroneous information that sometimes might be hurtful to individuals and to families. In Activity 3, students grapple with the difficulty of interpreting genomic information properly and consider the potential harm that can occur through misuse or abuse of personal genetic data.

Together, Activities 1-3 establish a context within which the *students themselves* should begin to ask questions about privacy and confidentiality. Your pedagogical goal during these activities is to foster their growing sensitivities to these issues and to help them build a foundation of understanding from which they will approach Activities 4 and 5, which address these questions specifically.

Related Information in the Teacher Narrative

You will find additional background information on the following pages of the teacher narrative:

□ content and use of registry databases (pp. 25–31)

□ ethical questions related to genetic registries (pp. 36–37)

Suggestions from the Field Test

□ Some students will not need the detailed instructions for pedigree construction that we provide on Worksheet 5, *Constructing a Pedigree*. Suggest that these students use the information they retrieve from the LGD to draw a pedigree for their "immediate family" and then construct the extended pedigree by talking to each other about how their individual pedigrees fit together.

- □ Remember that an important relationship exists between the number of students in a group and that group's ability to complete its family's pedigree. Groups that are too large may gather too much data to use effectively; groups that are too small will gather too little data to see the full pedigree. The guidelines provided on BLM T-2 will help you establish groups of the appropriate sizes. Be sure that students understand that they must share their data with other members of their group even if each student has access to his or her own computer.
- □ You may wish to suggest that students who finish their work early explore the LGD further and generate a list of the types of questions that someone using this database could answer.

Annotated Student Activity

Activity 1 assumes that students have read the background material we provide (Worksheet 3, *Genomic Databases*).

To activate the data required for this activity, select *Genetic Registries* from the *ACTIVITY* pull-down menu. You either may perform this function on each of your students' computers or ask them to complete it. The password is <Lyon>.

Imagine that you live sometime in the future, in a small town somewhere in rural America. Like other small towns (as well as big cities), your town finds itself increasingly affected by the rapid increase in genetic information that has resulted from the HGP. Members of the community, for example, can access a wealth of genetic data simply by searching the National Genome Database (NGD), a research database that is available online to anyone around the world who has the basic computer and telecommunications capabilities required to access it.

Imagine as well that about a year ago the public-health authorities decided to collect the growing body of personal, medical, and genetic data they had accumulated about the people of your community into a central database called the Local Genome-Database (LGD). Doctors, nurses, and public health officials can access this database to retrieve information important to their work.

Concerned about the public's response to this action, health authorities decided to make a small subset of these community health data available online. Their reasoning was that this would allow community members to learn about this resource and also give interested individuals a chance to explore the database and to provide input about the appropriate uses of these data. In the interest of community education, three families in the town volunteered to allow portions of their records to be made available to the public. These records include each family member's genotype for the following four genes: the fragile X gene (F); the gene for hereditary juvenile glaucoma (G); the angiotensinogen gene, a gene that can increase significantly an individual's risk for high blood pressure (A); and the gene for alpha hemoglobin (H).

Procedure

1. Your teacher will give you the name of the fictitious person whose identity you will assume for this activity. Your tasks are

- **□ to locate and examine your personal file in the LGD,**
- **□ to accumulate sufficient information to develop a pedigree that properly represents the structure of your extended family (your parents, siblings, aunts, uncles, cousins, and grandparents),**
- **□ to begin to interpret the genetic data you find in the LGD, and**
- **□ at the end of the activity, to decide whether you would like your fictitious person to be tested for his or her genetic profile for three additional genes.**

2. Take turns following the steps below to find the information that you need to complete Worksheet 4, *Collecting Family Data*.

Instruct the students to enter the appropriate information onto their worksheets as they locate it in the database. The students will need this information to complete their pedigrees.

Point out that the first column on their worksheet already has been completed for them. Although the person described may not be a member of their family and, if he is not, *should not* be used to generate their pedigree, his record is in the database. If students are confused about the type of information that they are looking for, suggest that they access Jamie's file and compare the information they find with the sample on the worksheet.

a. Locate your "personal file" in the LGD.

b. The screen that appears should be similar to the screen pictured in Figure 1. What types of information does it give you?

The screen shows the general information portion of the personal LGD record for one individual. The record lists the individual's sample number (a number arbitrarily assigned to an individual as his or her record was entered into the database), name, sex, age, and current status. It also lists the names of his or her immediate family members and provides some personal and family medical information.

c. Write your sample number and your fictitious name, age, parents' names, and sibling's or siblings' name(s) and sex(es) into the appropriate areas on your worksheet.

NOTE: Although each of you should complete your own computer search, you should work together to be sure that each member of your group retrieves the correct information. You will need this information to complete your family's pedigree.

Make sure that each student performs his or her *own* search. *Avoid allowing any group to designate one person to complete all of the required searches.* Remind your students that although each of them will record only his or her own information, they all will need information from each member of the group to complete their pedigree. Encourage the members of each team to work closely together to assure that the data that each person retrieves are accurate and complete.

Asking students to list their parents' and siblings' names and sexes will allow students to construct their family pedigrees in the manner in which geneticists usually do it, that is, by beginning with reported family relationships.

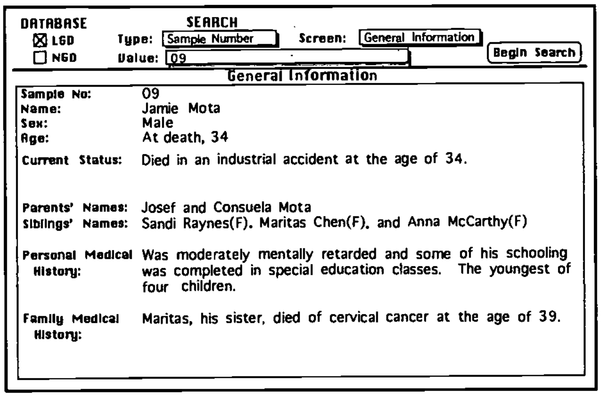
3. Use the data that you collected on Worksheet 4 to construct a pedigree for your extended family.

To Locate a Person's File in the LGD

- Click on the *LGD* checkbox.
- Click and pull on the *TYPE* box to highlight the line that reads *Name*. Release the mouse button.
- Click on the *VALUE* box. The computer will display a box that asks you to enter the name of the person for whom you wish to search.
- Enter your fictitious name (first and last) in the box and click on *OK* (*BEGIN SEARCH* in Windows). If there already is a name in the box, simply type over it.

NOTE: Remember that the computer searches for exact matches. Spaces count. If you make an error, press the [delete] key located on the upper right side of the keyboard. This will delete the last letter you entered.

TIP: You also may enter just the last name of your fictitious person and retrieve a list of everyone in the database with that last name. To retrieve the personal file of an individual whose name appears on the list, simply click on his or her name. To return to the list, click on the *RETURN TO LIST* button that will appear in the upper right of the screen. (**Windows Users:** Double click on a person's name on the list to retrieve his or her file. Click on *BEGIN SEARCH* to return to the list.)



General Information

Sample No: 09
Name: Jamie Mota
Sex: Male
Age: At death, 34

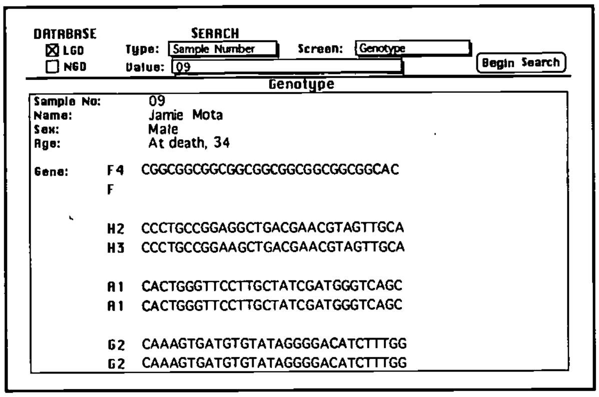
Current Status: Died in an industrial accident at the age of 34.

Parents' Names: Josef and Consuela Mota
Siblings' Names: Sandi Raynes(F). Maritas Chen(F). and Anna McCarthy(F)

Personal Medical History: Was moderately mentally retarded and some of his schooling was completed in special education classes. The youngest of four children.

Family Medical History: Maritas, his sister, died of cervical cancer at the age of 39.

Figure 1 Sample *General Information* screen from the LGD.



Genotype

Sample No: 09
Name: Jamie Mota
Sex: Male
Age: At death, 34

```
Gene:   F4  CGGCGGCGGCGGCGGCGGCGGCGGCGGCAC
        F

        H2  CCCTGCCGGAGGCTGACGAACGTAGTTGCA
        H3  CCCTGCCGGAAGCTGACGAACGTAGTTGCA

        A1  CACTGGGTTCCTTGCTATCGATGGGTCAGC
        A1  CACTGGGTTCCTTGCTATCGATGGGTCAGC

        G2  CAAAGTGATGTGTATAGGGGACATCTTTGG
        G2  CAAAGTGATGTGTATAGGGGACATCTTTGG
```

Figure 2 Sample *Genotype* screen from the LGD.

NOTE: You will find instructions for building your pedigree on Worksheet 5, *Constructing a Pedigree*. You will need the pedigree that you develop to complete Activity 2.

4. Take turns following the steps below to examine your personal LGD file more closely.

- **a. Again, locate your personal file in the LGD.**
- **b. Change to the *Genotype* screen to see your DNA profile for four different genes. The screen that appears should be similar to the screen pictured in Figure 2. What types of information does it give you?**

The screen shows seven or eight base sequences. Explain that these base sequences represent portions of the DNA from each of seven or eight chromosomes (one or two alleles for each of four genes, with symbols F, H, A, and G). Remind your students that humans are *diploid*, that is, that each of us normally inherits one full set of chromosomes from each of our parents. Consequently, we carry two copies of most genes, one copy maternally derived and one copy paternally derived.

To Change to the Genotype Screen

- Click and pull on the *SCREEN* box to highlight the line that reads *Genotype*.
- Release the button. The computer will display the new screen.

NOTE: When you first access a person's file, the computer always will display the *General Information* screen. To move back to this screen, click and pull on the *SCREEN* box, highlight *General Information*, and release the button.

The screen identifies these sequences as belonging to a particular individual by listing sample number, name, sex, and age. You will notice that the data also include allelic designations. That is, the symbols along the left side of this screen show both the letters that represent the various genes—F, H, A, and G—and the numbers—1, 2, 3, or 4—that indicate the particular allelic form of each gene that this person carries. *The capital letters do not indicate dominance.* To identify the genes to which these symbols refer and to discover which alleles are associated with traits that are expressed in a dominant fashion, the students will have to search the NGD. Students will do this in Activity 2.

Emphasize that the sequences shown are only 30 bases long and that this is much shorter than any known gene. The students should think of these sequences as *parts* of much longer sequences.

Some of your students still may be uncomfortable with the difference between a *gene* and an *allele*. Explain to these students that the symbols F, H, A, and G refer to different genes. These genes are associated with different traits and are located at different chromosomal positions. In contrast, the base sequences shown identify the particular forms—the particular alleles—of each gene that this person inherited from his or her mother and father. These sequences may match (that is, the person may be homozygous for that trait) or the sequences may differ (that is, the person may be heterozygous for that trait).

c. Discuss the information that you find on the *Genotype* screen and answer the questions on Worksheet 6, *Interpreting Data in the LGD*.

5. At the end of the period, your teacher will offer you an opportunity to have your fictitious person "tested" for his or her genotype for each of three additional genes (the genes associated with cystic fibrosis, sickle cell disease, and familial hypertrophic cardiomyopathy). Consider your decision carefully and follow your teacher's instructions for either consenting to the test or for refusing it.

Distribute three swabs and one envelope to each student and explain that if they were going to provide real samples for DNA analysis, they would use these swabs to collect cells from the insides of their cheeks. *Explain further that for reasons of hygiene and safety, they are not actually to swab the insides of their cheeks.* Instead, tell your students to place their unused swabs inside their envelopes and prepare the envelopes as described below.

- □ Write their own name and the name of their fictitious person on the front of the envelope.
- □ Write the word "yes" on the envelope if they want their fictitious person to be tested, or "no," if they have decided to decline the offer. Emphasize that each student must respond to the invitation in writing by either accepting the testing ("yes") or declining it ("no"). Emphasize as well that the lab will test either for all three genes or for none (that is, students may not ask to be tested for one gene but not the others).
- □ Hand in their *unsealed* envelope. Students should return their labeled, unsealed envelopes whether or not they choose to have their person tested.

Although we recommend that you answer the questions students may ask about the testing, we strongly suggest that you *do not volunteer* any information that they do not ask for specifically. For example, if the students ask what will happen to the results of the tests, answer that the results will be available on the day that the class completes Activity 4 and that the class will decide together how to handle them. Likewise, if the students ask for more information about the genes involved, briefly describe the trait with which each is associated. (You will find descriptions of these genes in the NGD and the *Glossary*.) If the students do not ask about these issues, do not raise them yourself. The students will discover how important these questions are as they complete Activity 4.

HOMEWORK ASSIGNMENT

Read the background information for Activity 2 (Worksheet 7, *Kate and Ryan*) and answer the study questions provided.

ANNOTATIONS TO WORKSHEET 6
Interpreting Data in the LGD

1. **Notice that your person's record contains a list of base sequences for genes F, H, A, and G. Why does this record show two sequences for most or all of these genes?**

Humans are *diploid*, that is, each of us normally inherits one full set of chromosomes from each of our parents. Consequently, we carry two copies of most genes, one copy maternally derived and one copy paternally derived.

2. **The sequences shown are only 30 bases long. Do you think that each sequence represents one complete gene or only a portion of a gene? Explain your answer.**

Each sequence probably represents only a portion of one gene. The average human gene is about 20,000 bases.

3. **What explanation can you offer for the observation that males have only one F gene and never two?**

The F gene likely is located on the X chromosome. Females inherit two X chromosomes, one from each parent. In contrast, males inherit only one X, from their mothers.

4. **Consider genes H, A, and G. For which of these genes are you homozygous? For which are you heterozygous? List two ways in which you can determine this from the information on the screen.**

Answers will vary. Students can determine homozygosity or heterozygosity for these genes by comparing either the gene symbols or the actual base sequences of each pair of alleles.

5. **Notice that the LGD does not give you any information about the genes symbolized by these letters. What type of database would you have to search to find information about these genes?**

Students would have to consult a research database to retrieve information about these genes.

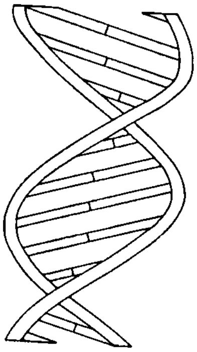
ANNOTATIONS TO WORKSHEET 7
Kate and Ryan

1. **Why did Stanley think there had been a mixup in processing Kate's sample? Use information provided in the pedigree on p. S-15 in your answer.**

Stanley recognized that for Chris to be dF508/dF508, Chris's mother, Kate, must carry at least one dF508 allele. If Kate is really Chris's mother, her genotype cannot be N/N.

2. **Suppose Kate's test had shown her to have the genotype dF508/dF508. Do you think Stanley would have questioned this result? Explain.**

Yes, again Stanley might have questioned the result. Although Kate's genotype then would be compatible with Chris's (that is, she would carry the dF508 allele), in this case, her phenotype would not match her genotype (dF508/dF508).



Activity 2
Explaining the Outliers

Focus

In Activity 1, the students used the data that they retrieved from the LGD to create pedigrees for their extended families. In Activity 2, the students verify these pedigrees against those stored in the database and discover that the LGD contains data that are contradictory. Using both the LGD and the NGD (the model research database) as sources of information, the students work together to develop and test a set of hypotheses about the reasons for these contradictions. The searches provide students with an opportunity to explore a research database and lead students to consider several issues of importance to the HGP, including the relationship between genotype and phenotype, the familial nature of genetic data, and the fact that databases can contain errors of various types.

HGP Context

The HGP is focused primarily on constructing communal research databases that eventually will contain a full, anonymous description of the human genome. Simultaneously, however, scientists around the world (many of whom are associated with the HGP and many of whom are not) also are constructing a large number of additional research databases that contain a variety of other types of information related to the human genome. These databases may contain specialized clinical information that might be helpful in the diagnosis of genetic disorders, information about the probes and other biological materials scientists use to identify genes, pedigree data that describe families that carry various genes of interest, or even enormous volumes of physical mapping data that describe thousands of cloned DNA fragments. Some of these databases are widely accessible; others, such as those that store physical mapping data, often are developed, maintained, and used entirely within a single laboratory.

Likewise, many hospitals, insurance companies, employers, public-health agencies, and schools maintain registry databases that contain personal, health-related information about specific individuals and about groups of individuals. Most of these databases do not contain genomic information at the level of DNA sequences. In this sense, the LGD associated with this module is "futuristic." Nevertheless, our increasing awareness of the role of heredity in personal health and our increasing ability to identify the presence of specific genes through laboratory testing suggest that the volume and the level of detail of the personal genetic information that such agencies collect and store in registries inevitably will increase.

Researchers and health-care workers often move from one type of database to another in the course of their work. GDB, GenBank, and OMIM, for example, each contains some portion of the total information currently known about the fragile X gene. An investigator interested in compiling this information would have to consult GDB to find detailed mapping data, GenBank to retrieve the sequence data, and OMIM to obtain a description of the phenotypes associated with the condition. A scientist interested in studying the structure of the DNA in the area of the fragile X gene also might search ATCC to identify the probes or cell lines available to aid her in her research.

As different types of data become available, the demand for easy access to the whole range of genetic information, wherever it is stored, likely will grow. Research scientists and health-care professionals find convenient, rapid access essential to their work when information changes daily and the amount of managed data already is beyond the ability of print-based media to contain it. In fact, one of the goals of the informatics component of the HGP is to use telecommunications to link many of these databases, creating, in effect, a single genomic database through which users will complete all of their genome-related searches.

In Activity 2, the students will see how one can move effectively from one type of database to another to answer questions of interest. They will see how one can learn quickly about the genetic profile of an entire family, but also how unexpected data can appear in unexpected ways. Students will have an opportunity to learn about how such unanticipated results can arise, the sensitive issues they may raise in families, and the need for accurate sample acquisition and data entry. These experiences establish an important foundation for subsequent activities having to do with the ethical issues of privacy and confidentiality and the public-policy dilemmas that these issues create.

Major Concepts

□ In the absence of mutation from one generation to the next, children carry alleles representative of their biologic parents. The discovery of a genotypic incompatibility within a family may indicate either that the genetic relationships within the family are not as reported or that an error occurred in the collection, reporting, or interpretion of the data.

□ A hypothesis is a proposed explanation for a question, problem, or observation. The most useful hypotheses are testable.

□ We should not assume that data always are error free. Errors that appear in data that are stored in electronic databases may result from a number of problems, including sampling error, data-entry error, or malicious mischief. Retesting the individuals whose data are in question may correct such errors.

Student Objectives

As students complete this activity, they should

□ understand the differences between the types of information that are stored in registries and research databases,

□ search a registry database for genomic data,

□ search a research database for information that explains the personal data in the registry,

□ understand the logic involved in developing and verifying pedigrees based on reported family relationships and genotypic data,

□ identify individuals in a fictitious pedigree whose genotypes are incompatible with their reported family relationships,

□ develop a set of hypotheses to explain these contradictory data,

□ search both a research database and a registry to retrieve information that would be useful in distinguishing among alternative hypotheses, and

□ recognize that databases may contain errors, but that retesting allows us to correct some of those errors.

Science Process Skills

□ Comparing

□ Generating hypotheses

□ Gathering data

□ Analyzing data

□ Drawing conclusions based on the results of research

□ Using the computer as a research tool

Materials for a Class of 30

□ 30 sets of student text, pp. S-17–S-18. Because students will complete their written work on their own paper or on the worksheets provided, other students can reuse these pages.

□ 30 copies of Worksheet 8, *Analyzing the Discrepant Data.*

Advance Preparation

□ Make the necessary student copies.

□ Homework: Ask your students to read the background information for Activity 2 (Worksheet 7, *Kate and Ryan,* pp. S-15–S-16) and answer the study questions provided. (Answers are on p. 82.)

□ Set the software on each computer to *Explaining the Outliers, Part A.* You will find instructions for how to do this on p. 56. The password is <Sutton>. You may wish to have your students activate the appropriate data; we provide a blackline master listing the steps required at the end of Activity 1 (BLM T-3).

Estimated Time

Two 45-minute class periods

Introduction

The HGP is likely to have a tremendous impact, not only on biology, but on many other areas of life as well. Scientists expect that many new insights will emerge as we locate and identify genes, as we gain new understandings about how genes are organized and regulated, and as we are able to compare the human genome directly with the genomes of other organisms. Likewise, although finding every human gene will not assure a cure for all genetically based disease, it will allow us to ask increasingly sophisticated questions about the biochemistry underlying genetic disorders and to develop increasingly sensitive methods for predicting, diagnosing, and treating them.

Although our growing understanding of the relationship between genes and human health will bring with it many benefits, it also will raise some serious issues. Not surprisingly, some of these issues relate to decisions that individuals make about themselves and their families. For example, the ability to predict the likelihood of future disease based on our individual or family genetic profiles may influence decisions that we make about our life styles and about employment, marriage, and reproduction. Other issues, however, relate to how various private and government organizations may seek to use genetic data. For example, the predictive power of increased genetic knowledge also may trigger changes in the health-care industry and in the policies and practices of health and life insurance companies.

Although these issues are not the primary focus of Activity 2, you can use the activity to begin to build a context within which—in Activities 4 and 5—you will discuss some of these questions with your students. Activity 1 required the students to create pedigrees for the extended families whose genetic profiles are stored in the LGD. As students compare the pedigrees that they generated with the pedigrees stored in the database, point out to them that their genetic profile is not entirely their own; rather, it is connected to, and shared with, the profile of every member of their family, both immediate and extended. Encourage the students to think about how this may affect the ways in which individuals choose to share and to use such information and how it may affect decisions that we make as a society about allowing the collection of such information.

Likewise, when the students compare their pedigrees with those stored in the database, they will discover one individual in each family whose genotype is incompatible with his or her reported position in the family. As each group works to determine the reasons for the contradictions within the registry data, ask the students to think about the possible implications of these contradictions and point out that new technologies bring with them many benefits, but also many unexpected consequences. Remind them that they should recognize the possibility that even electronically stored and disseminated data may contain errors.

Finally, as the students finish their work and give their group reports, ask them to consider what their experiences suggest about the benefits as well as the risks of collecting, storing, and *using* genetic data in ways that may affect people's lives.

Related Information in the Teacher Narrative

You will find additional background information in the following pages of the teacher narrative:

□ genetic data as family data (pp. 36–40)

Annotated Student Activity

You may wish to begin class with a brief review of the Kate and Ryan Dozark case. It is important that students understand the genetic incompatibility involved and the steps by which the discrepancy was resolved.

One of the goals of the introductory activity was to give you some sense of the nature and the volume of genomic information that researchers associated with the HGP are generating. Without computers, these data already would be unmanageable.

In Activity 1, you used the information that you retrieved from the LGD to construct a pedigree for your extended family. Scientists and health-care professionals use pedigree data in many ways. For example, investigators interested in identifying relationships between genes and specific human traits (including disease conditions) may use pedigree data to help establish the heritable nature of a trait and to determine the specific pattern of inheritance involved. Likewise, genetic counselors and other health-care professionals may construct pedigrees to help explain the principles of inheritance to members of families who carry genes associated with known diseases. The possibility exists that some day public-health officials also may use pedigree data to trace individuals at risk for certain preventable genetic conditions.

In this activity, you will have the opportunity to investigate the data in the LGD more thoroughly. You also will access a research database, the National Genome Database (NGD), for information that will help you better understand the data in the Local Genome Database (LGD).

To Access the Pedigree for Your Extended Family

- Click on the *LGD* checkbox.
- Click and pull on *TYPE* to highlight the line that reads *Name*. Release the mouse button.
- Click on *VALUE*. Enter the name of a family member and click on *OK* (or *BEGIN SEARCH*).
- Once the *General Information* screen appears, click and pull on *SCREEN* and highlight *Pedigree*. Release the mouse button and the computer will display the pedigree for the appropriate extended family.

PROCEDURE

1. **Your primary task in Activity 1 was to determine, on the basis of a rather small amount of data, the structure of the pedigree for your extended family. Follow your teacher's instructions for forming the same work groups in which you participated during Activity 1.**
2. **Check your pedigree against the pedigree for your family that was generated by the computer and stored in the LGD.**

As explained in the implementation support section of the module (p. 54), students are unable to access the computer-based pedigrees until you activate the data for *Explaining the Outliers, Part A*. You should have completed this step on each computer before beginning Activity 2. You will find instructions for doing so on p. 56. The password is <Sutton>.

Special Instructions for the Pedigree Screens

The computer-generated pedigree identifies each individual by sample number rather than by name. To see an individual's name, click and hold on the appropriate pedigree symbol. A pop-up box will appear that lists this information, as well as the individual's sample number, sex, age, and genotype.

If you move the cursor off the symbol while holding the mouse button, the box will remain visible on the screen after you release the button. Click on the box to make it disappear.

(**Macintosh Users:** If you wish to go immediately to a specific person's record without going through a search, press the [option] key while you click the mouse on the symbol for the person whose data you wish to see. The database will show you the *General Information* screen for that person. To return to the *Pedigree* screen again, simply click on the *SCREEN* box, highlight *Pedigree*, and release the mouse button.)

3. Compare the pedigree that appears on the screen to the one that you assembled.

a. Are there any differences? If there are, discuss the possible reasons for these differences with the other members of your group.

NOTE: A diagonal line through a symbol indicates that this person is deceased. It is possible that the pedigree that you drew does not show this detail. Although you may wish to add this information to your pedigree, its absence will not affect your ability to complete the activity.

Because the students have assembled their pedigrees from reported family relationships, it is unlikely that they will have made errors. This means that with the exception of the outliers (individuals whose genotypes do not fit with those of the rest of the family) described below, the students' pedigrees should match those they retrieve from the LGD. Because the students were not given all people's current ages, their pedigrees may show siblings out of proper birth order.

As described in the introduction to this activity, each of the computer-based pedigrees reveals that there is one individual whose relationship to the family is in question. The symbol for this outlier is indicated with an asterisk and is not connected to the pedigree. The designation "Incompatible Genotype" appears in the associated pop-up box.

b. The pedigree on the screen reveals an individual who is not connected to the pedigree and is indicated by an asterisk. This person is the "outlier" referred to in the title of this activity. Discuss the possible reasons for this person being classified as an "outlier."

An outlier is a person whose genotype does not fit with those of the rest of the family. The use of the term "outlier" indicates that this individual's genotype lies "outside" the range of genotypes that one would predict for his or her position in the family.

4. What does the phrase "Incompatible Genotype" mean? Did you take genotypes into consideration when you drew your pedigree? If not, on what basis did you construct your pedigree?

The phrase "Incompatible Genotype" suggests that one or more of this person's alleles could not have been inherited from his or her mother, father or, perhaps, from both. The students constructed their pedigrees on the basis of *reported* family relationships. This is, in fact, how most genetic analyses of families begin. If the students had used genotypic data to build or to confirm their pedigrees, they would have discovered the discrepancies themselves.

5. Complete Steps I through V on Worksheet 8, *Analyzing the Discrepant Data*, to develop and test some hypotheses about the reason for the apparent discrepancy between the family data and the genotypic data in the LGD. Record your answers in the appropriate places on the worksheet.

Distribute one copy of the worksheet to each student. At this time, also set the software on each computer to *Explaining the Outliers, Part B*. This change in the software activates the retest function that students will need to use to test their hypotheses about the discrepant data. The password is <Avery>.

6. Follow your teacher's instructions for reporting the results of your investigation to the class.

To conclude the activity, ask each group to summarize briefly the results of its work on its family's pedigree. Ask the groups to focus on

□ the nature of the discrepancy,
□ the possible hypotheses to explain the discrepancy, and
□ their conclusion and reasons for reaching it.

It is important that your students understand that the point of this activity is not to identify the "correct" explanations as such, but to understand that discrepancies can occur in data, to think about possible reasons for such discrepancies, and to see that systematic analysis can yield viable explanations for data that initially may be confusing. Emphasize the importance of retesting to confirm questionable information and remind the students that retesting provided a straightforward way to resolve one of the three discrepancies in the LGD (Drew Schmidt).

If your students are uncomfortable with the tentative nature of their conclusions, note that we have structured the module so that students eventually can discover very plausible explanations for two of the three discrepancies. Students identify the laboratory error in this activity (Jacobs family), and, during Activity 3, encounter genetic anticipation and the expansion of trinucleotide repeats as an explanation for the discrepancy in the Mota family. The incompatibility in the Thomas family is the only case left ambiguous, but because of the extensive mismatch between Lisa's genotype and those of her parents, the most likely explanation for this discrepancy is adoption.

ANNOTATIONS TO WORKSHEET 8
Analyzing the Discrepant Data

I. Describe the problem.

A. One of the pedigrees in Figure 5 illustrates the portion of the pedigree that is at issue in your family. Examine the pedigree that the computer provides for your family and write the full genotype for each of the individuals involved below his or her name. One genotype has been written in for you already.

Be sure that the students copy each person's *full* genotype onto their worksheets, using the correct symbols. Explain that they need only to complete the portion of Figure 5 (a, b, or c) that applies to *their* fictitious family.

Remind the class that they can retrieve an individual's name and genotype by clicking on the appropriate symbol on the pedigree and then moving the cursor outside the box. A pop-up box will appear that lists this information. If they click on the pop-up box again, the box will disappear.

B. 1. Is there an incompatibility in the genotypes that you have listed? Explain.

NOTE: In each case, the genotypes are given so that the gene a person inherited from his or her mother is listed first. For example, a genotype of A1/A2 indicates that the person inherited an A1 allele from his or her mother and an A2 allele from his or her father. When the alleles in question might have been inherited from either parent, the allele symbols are listed in parentheses. Take this into consideration when you determine the nature of the genetic incompatibility that is involved in your fictitious family.

2. If there *is* an incompatibility, does it involve only one gene, or does it involve more than one gene?

3. Does it involve one parent or both parents? Be specific.

Explain to the students that to answer these questions properly, they will have to examine the complete genotypes of at least three individuals (the person identified as having an incompatible genotype, his or her father, and his or her mother). If necessary, also explain to them that in the absence of mutation from one generation to the next, children carry only those alleles that are represented in their biologic parents. (As students will see in Activity 3, a very special type of mutational event is the explanation for the discrepancy in Walter's case.) Finally, make sure that students understand that the genotypes are listed with the maternally derived allele *first*. If students do not understand this, they may fail to see the discrepancy in the data they have been given.

WALTER:

1. Yes, there is an incompatibility.
2. The incompatibility involves only the "F" gene.
3. For this gene, Walter's genotype is consistent with his father's genotype (Walter inherited a Y chromosome from his father), but not with his mother's (Walter carries an F4 allele, whereas his mother carries the genotype F1/F3).

DREW:

1. Yes, there is an incompatibility.
2. The incompatibility involves only the "H" gene.

3. For this gene, Drew's genotype is consistent with his father's genotype (Drew could have inherited an H4 from his father), but not with his mother's (Drew carries two H4 alleles, whereas his mother carries the genotype H1/H1).

LISA:

1. Yes, there is an incompatibility.
2. The incompatibility involves three of the four genes (the F, H, and A genes).
3. It involves both parents. Lisa's genotype is incompatible with her father's genotype with respect to the F gene, and her genotype is incompatible with her mother's genotype with respect to the H gene. Her genotype is incompatible with both parents' genotypes with respect to the A gene.

II. Collect information that may help you solve the problem.

Your work in Step I should have convinced you that the data in the LGD contain an important discrepancy. Although reported family relationships indicate that this individual is a member of this family, his or her genotype is incompatible.

An important step in solving any problem is to collect relevant information. In this case, for example, it may be important to know something about the genes that are displayed on the pedigree and to understand how the related phenotypes are or are not expressed in the individuals involved.

Take turns following Steps A and B to collect information about the genes represented in Drew's, Walter's, or Lisa's file.

A. Choose one of the genes to study in more detail. Search the NGD (the model research database) for information about the gene you selected. Enter the information that you retrieve into Table 2.

NOTE: Each of you should choose a different gene to investigate.

Symbol	Gene Name	Definition	Symbols for the Known Allelic Variations and Their Associated Phenotypes
F	fragile X mental retardation	gene associated with the most frequently encountered form of inherited mental retardation in humans	F1 - No information is available about the phenotypic significance of this allele. F2 - No information is available about the phenotypic significance of this allele. F3 - No information is available about the phenotypic significance of this allele. F4 - No information is available about the phenotypic significance of this allele.
H	alpha hemoglobin	gene for the alpha subunit of the hemoglobin molecule	H1 - normal hemoglobin H2 - normal hemoglobin H3 - normal hemoglobin H4 - normal hemoglobin
A	angiotensinogen	gene for the polypeptide angiotensinogen. One allele is associated with a predisposition to high blood pressure.	A1 - normal A2 - predisposition to high blood pressure
G	glaucoma	gene associated with hereditary juvenile glaucoma	G1 - increased risk of glaucoma G2 - no increased risk of glaucoma

Table 2 Information about the genes represented in Drew's, Walter's, and Lisa's file. [Answers will vary. We have entered sample answers for one team.]

To Locate the File on a Gene in the NGD

- Click on the *NGD* checkbox.
- Click and pull on the *TYPE* box and highlight the line that reads *Gene Symbol*. Release the mouse button.
- Click on the *VALUE* box. The computer will display a list of symbols. Highlight the symbol of the gene for which you wish to search and release the mouse button.
- Click *BEGIN SEARCH* to start the search.
- If necessary, use the scroll bar to read all of the notes.

To Change to the Allelic Variations Screen

- Click and pull on the *SCREEN* box to highlight the line that reads *Allelic Variations*.
- Release the button. The computer will display the new screen.

NOTE: When you first access a gene's file, the computer always will display the *General Information* screen. To move back to this screen, just click and pull on the *SCREEN* box, highlight *General Information*, and release the button.

B. What is Drew's, Walter's, or Lisa's genotype? Based on the information your team has collected about the genes involved, does his or her genotype match his or her phenotype? Explain.

WALTER:
Walter's genotype is F4, H2/H4, A1/A1, G2/G1. To the extent that it can be determined, Walter's phenotype matches his genotype. A key observation is that Walter carries the allele (G1) for hereditary juvenile glaucoma and already shows the associated phenotype. This form of glaucoma is relatively rare in this population. The NGD shows that no information is available about the F4 allele, so we cannot determine whether his learning difficulties are consistent with his genotype for the fragile X gene.

DREW:
Drew's genotype is F1, H4/H4, A1/A2, G2/G2. To the extent that it can be determined, Drew's phenotype matches his genotype. A key observation is that Drew carries the angiotensinogen allele (A2) associated with a predisposition to high blood pressure and already shows the related phenotype.

LISA:
Lisa's genotype is F3/F2, H4/H4, A1/A2, G2/G2. To the extent that it can be determined, Lisa's phenotype matches her genotype. A key observation is that Lisa carries the angiotensinogen allele (A2) associated with a predisposition to high blood pressure and already shows the related phenotype.

III. Develop hypotheses.

What hypotheses can you propose to explain the discrepant information in the LGD? List your hypotheses below.

You may have to help the class generate some hypotheses for the incompatible genotypes that the students have found. Some of the possible hypotheses include the following:

□ a child who was adopted or who is living with a stepmother or stepfather after a remarriage.

□ someone whose parent is not as reported. This may be a case of nonpaternity, where the putative (reported) father is not the biological father because of artificial insemination using donor sperm. Or, in other cases more difficult to deal with, the biological father may be known only to the mother. In this era of *in vitro* fertilization, a nonmaternal pregnancy also is possible if a donor egg were used.

□ sample errors. On rare occasions, laboratory personnel might mix up samples between the time they take them (as blood or hair) and the time they analyze them. This may be more likely when laboratories obtain samples from people with similar names at the same time.

□ data errors. Errors can occur as a result of mistakes during the sequencing process, software problems, data-entry mistakes, and even through disc failure or computer malfunction. Although the increasingly common practice of transmitting data directly from one computer to another (for example, from analysis in one computer to storage in another) is reducing data-entry errors, data errors still occur.

□ malicious mischief. Unethical individuals intentionally may manipulate samples, data, or computer files so that the database contains false data. Administrators, laboratory personnel, and information specialists who are responsible for the integrity of the data use passwords and limited-access files to prevent such tampering, but it still may occur.

□ mutation. Changes can occur in the DNA from one generation to the next. Such changes may alter the particular allelic form that an individual receives for a gene.

IV. Test your hypotheses.

A. Two hypotheses that could explain the discrepancy are sample errors (for example, a lab technician may have confused this person's sample with someone else's) or data error (for example, this person's sequence for this gene might have been entered incorrectly or a lab technician may have made an error as the DNA was sequenced). How could you determine whether there has been a mixup in one of these steps?

To Retest the DNA Profile for an Individual in the LGD

- Activate the retest function by clicking and pulling on *ACTIVITY* at the upper left of your screen and highlighting the line *Explaining the Outliers, Part B.*
- Your teacher will give you the password to enter. Click on *OK.*
- Select the LGD database.
- Return to the *Pedigree* screen for your family by doing a name search and highlighting the *Pedigree* screen. A button labeled *RETEST* will appear. (The *RETEST* button will appear only on the *Pedigree* screen.)
- Click on *RETEST* and enter the first and last name of the individual whom you wish to retest. Click on *OK* to complete the retest.
- You will see a message reporting the results of the retest. Read the message, then click on *OK* to continue your work.

NOTE: Because of the relatively high cost of DNA testing, the lab will accept only four retest requests from you during this work session. Choose the individuals you wish to retest carefully.

The most direct way to evaluate these hypotheses is to retest the DNA of the individuals in question. The retest function is available in *Explaining the Outliers, Part B.* The password is <Avery>.

1. Whom did you choose to retest, and what were the results?

Person(s) retested [Answers will vary.]

Results [Answers will vary.]

2. Has retesting resolved the discrepancy? Explain. [Answers will vary.]

WALTER and LISA:
No, retesting simply confirms the data already in the database. This suggests that neither sample error nor data-entry error is the reason for the observed discrepancy.

DREW:
Retesting should have resolved the discrepancy. If the students retest Paul, Peggy, and Drew Schmidt, they will discover that the retest corrects errors in both Paul's and Drew's files. Paul *had* been listed as H4/H1; the retest corrects his genotype to H4/H4. Drew *had* been listed as H4/H4; the retest corrects his genotype to H1/H4.

Note that one plausible explanation for these errors is that Paul's and Drew's samples or data were switched by mistake sometime during sample processing or data entry. The fact that Drew's legal name is Paul Andrew Schmidt, like his father's, lends additional credibility to this explanation.

Encourage your students to watch their screens closely during the retest process. When the computer adds the retest results on Drew to the database, his corrected genotype will appear beneath his symbol and a new line will appear connecting Drew to his family's pedigree.

NOTE: The genotypic incompatibility between Drew and his mother disappears as soon as Drew is retested and the new data are entered into the database. If the students test Drew before they test Paul, they may be satisfied with seeing the change in Drew's file and they may not probe further. Encourage the students to continue their retesting, even after Drew's data are corrected.

You might point out, for example, that the change in Drew's file suggests that a sampling error may have occurred. Suggest that they test further to see whether they can determine how that error might have happened. Careful students may go on to test Paul, and will discover that it looks as though Paul's and Drew's samples may have been exchanged by mistake.

B. Some of your hypotheses may have stated that Drew, Walter, or Lisa is not the genetic child of one or both of their parents.

1. In theory, what types of evidence available from the LGD would support a hypothesis of adoption?

2. What types of evidence would support nonpaternity?

3. What types of evidence would support nonmaternity?

You may have to help students with this reasoning.

Hypothesis	Evidence That Would Support the Hypothesis
1. adoption	The incompatibility might involve more than one gene and might involve both parents.
2. nonpaternity	The individual's genotype would match his or her mother's genotype but might not match the father's. The incompatibility might involve more than one gene.
3. nonmaternity	The individual's genotype would match his or her father's genotype but might not match the mother's. The incompatibility might involve more than one gene.

NOTE: Although retesting should have eliminated the discrepancy in Drew's family, your students still should complete the worksheet. This will assure that they will be able to understand the reports that their classmates make on the other two families.

C. How does Drew's, Walter's, or Lisa's genotype fit with those of his or her parents? Explain.

Walter's genotype is incompatible with his mother's and compatible with his father's.

Drew's genotype was incompatible with his mother's and compatible with his father's.

Lisa's genotype is incompatible with both her mother's and her father's. The incompatibility also involves several genes.

V. State your conclusion.

A. What seems to be the simplest explanation for the observed incompatibility in your fictitious family?

WALTER:
Probably the simplest explanations are nonmaternity, adoption, or mutation. Because Walter's genotype matches his phenotype (to the extent to which we can know this), it seems unlikely that the incompatibility is the result of sample or data error. The results of retests on Walter and his parents also confirm the data currently in the database. Walter's incompatibility with his mother's genotype suggests either that Walter is adopted (it is possible that Walter's compatibility with his father's genotype is coincidental) or, more likely, that Walter is Mark's genetic son, but not Sandi's. Sandi may be Walter's stepmother; alternatively, Sandi may have received a donor egg. (The language "genetic son" rather than "biologic son" distinguishes the case in which a woman gives birth to a child conceived from her own egg from the case in which a woman gives birth to a child conceived from another woman's egg).

The actual explanation for Walter's incompatibility will become clear in Activity 3 with the explanation of the phenomenon of genetic anticipation. Walter could have an F4 allele and still be Sandi's genetic son.

LISA:
Lisa is most likely *not* the genetic child of either of her parents.

DREW:
Probably the simplest explanation is that Drew's

sample and his father's sample were mixed up during the processes of sampling, testing, or data entry. This would suggest that the genotype originally listed in the LGD for Drew was really that of his father, and vice versa. Note that Drew's legal name is Paul Schmidt, like his father's. Technicians might have confused the two samples involved if the samples became separated from the accompanying paperwork.

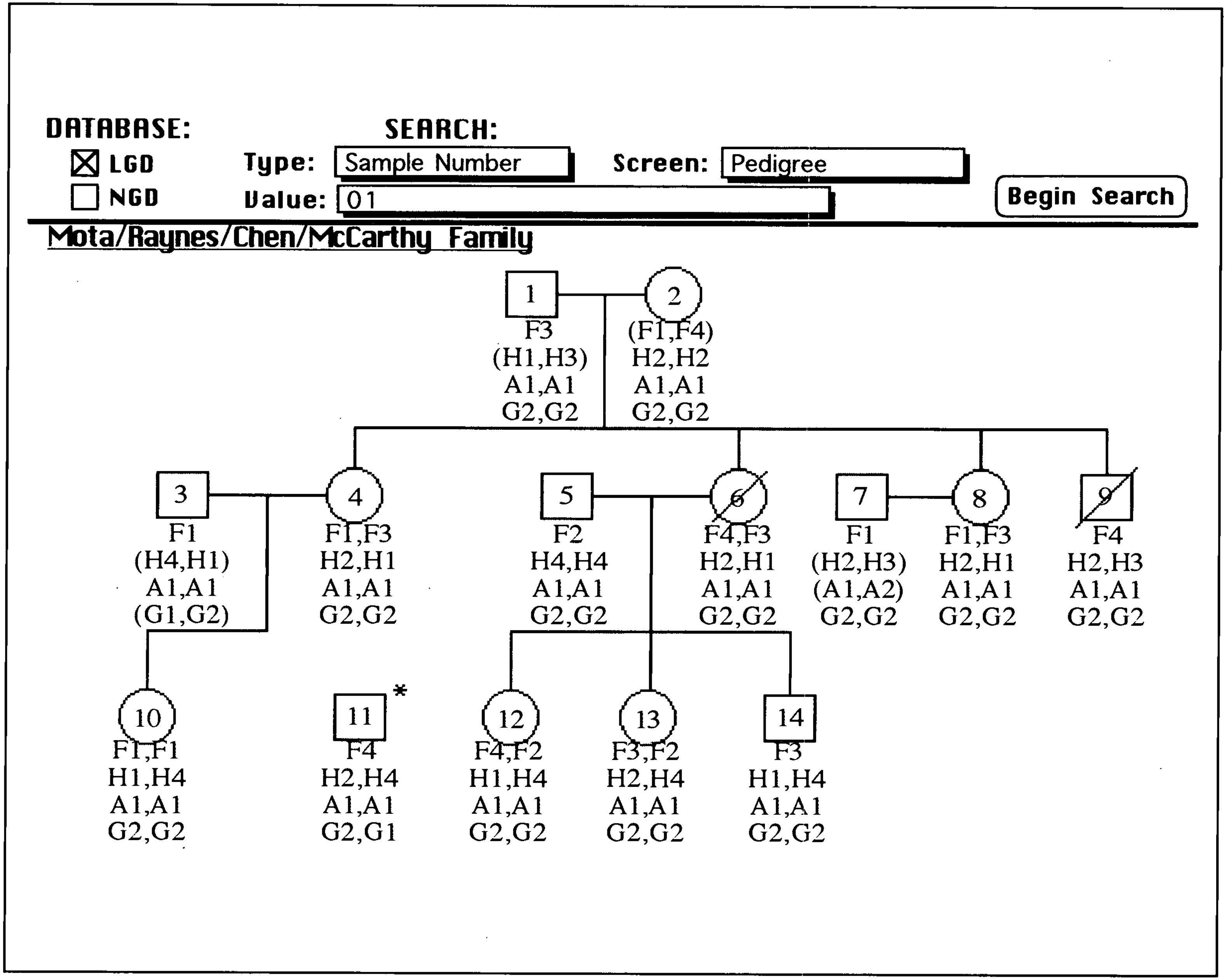
B. How strong is the evidence for your conclusion? Explain.

WALTER:
At this point, the conclusion is tentative. Students cannot eliminate any of the possible hypotheses on the basis of the data available in the LGD. Students should understand that this sometimes is the case when dealing with questions of this sort.

LISA:
At this point, the conclusion is tentative. Students cannot eliminate any of the possible hypotheses on the basis of the data available in the LGD. Students should understand that this sometimes is the case when dealing with questions of this sort.

You might point out that if Lisa is adopted, this discrepancy actually represents an error in the database. Emphasize the fact that if relationships are not as they are reported to be, this interferes with the usefulness of the data in the database. One misattributed relationship, for example, can seriously compromise a study of transmission patterns.

DREW:
The conclusion is well supported. Retesting both Paul and Drew reveals the errors in the original laboratory results. Drew's corrected genotype is compatible with his father's corrected genotype and with his mother's genotype.

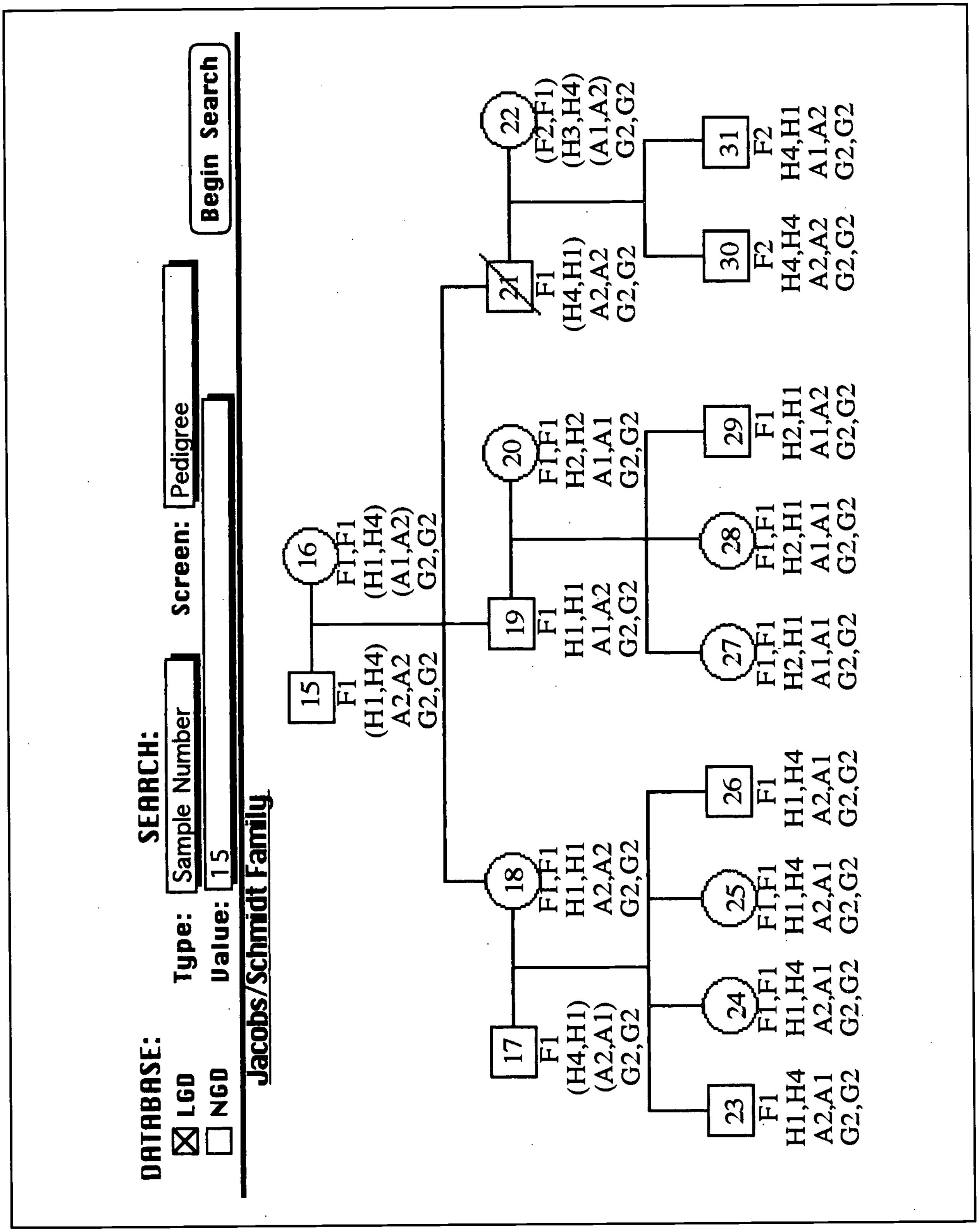


DATABASE:
☒ LGD
□ NGD

SEARCH:
Type: Sample Number
Value: 01

Screen: Pedigree

Figure T-1 Pedigree screen for the Mota family. Note that maternally derived alleles are listed first and paternally derived alleles are listed second. Parentheses indicate cases in which maternal or paternal inheritance cannot be determined.



DATABASE: ☒ LGD □ NGD

SEARCH: Type: Sample Number Value: 15

Screen: Pedigree

Jacobs/Schmidt Family

15 F1 (H1,H4) A2,A2 G2,G2

16 F1,F1 (H1,H4) (A1,A2) G2,G2

17 F1 (H4,H1) (A2,A1) G2,G2

18 F1,F1 H1,H1 A2,A2 G2,G2

19 F1 H1,H1 A1,A2 G2,G2

20 F1,F1 H2,H2 A1,A1 G2,G2

21 F1 (H4,H1) A2,A2 G2,G2

22 (F2,F1) (H3,H4) (A1,A2) G2,G2

23 F1 H1,H4 A2,A1 G2,G2

24 F1,F1 H1,H4 A2,A1 G2,G2

25 F1,F1 H1,H4 A2,A1 G2,G2

26 F1 H1,H4 A2,A1 G2,G2

27 F1,F1 H2,H1 A1,A1 G2,G2

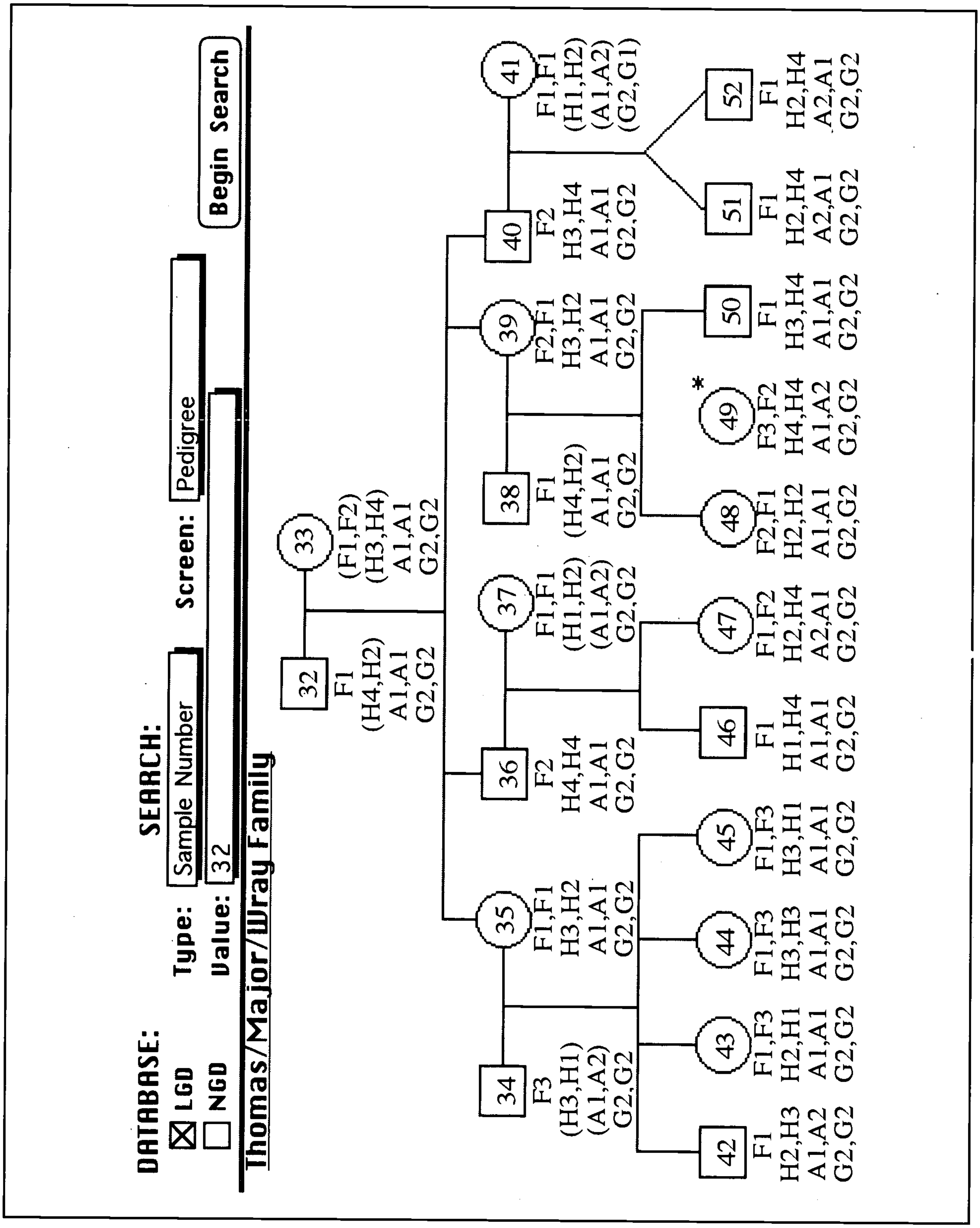
28 F1,F1 H2,H1 A1,A1 G2,G2

29 F1 H2,H1 A1,A2 G2,G2

30 F2 H4,H4 A2,A2 G2,G2

31 F2 H4,H1 A1,A2 G2,G2

Figure T-2 Pedigree screen for the Jacobs family. Note that maternally derived alleles are listed first and paternally derived alleles are listed second. Parentheses indicate cases in which maternal or paternal inheritance cannot be determined.



DATABASE: ☒ LGD □ NGD

SEARCH: Type: Sample Number Value: 32

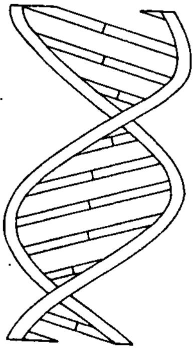
Screen: Pedigree

Begin Search

Thomas/Major/Wray Family

32 F1 (H4,H2) A1,A1 G2,G2
33 (F1,F2) (H3,H4) A1,A1 G2,G2
34 F3 (H3,H1) (A1,A2) G2,G2
35 F1,F1 H3,H2 A1,A1 G2,G2
36 F2 H4,H4 A1,A1 G2,G2
37 F1,F1 (H1,H2) (A1,A2) G2,G2
38 F1 (H4,H2) A1,A1 G2,G2
39 F2,F1 H3,H2 A1,A1 G2,G2
40 F2 H3,H4 A1,A1 G2,G2
41 F1,F1 (H1,H2) (A1,A2) (G2,G1)
42 F1 H2,H3 A1,A2 G2,G2
43 F1,F3 H2,H1 A1,A1 G2,G2
44 F1,F3 H3,H3 A1,A1 G2,G2
45 F1,F3 H3,H1 A1,A1 G2,G2
46 F1 H1,H4 A1,A1 G2,G2
47 F1,F2 H2,H4 A2,A1 G2,G2
48 F2,F1 H2,H2 A1,A1 G2,G2
*
49 F3,F2 H4,H4 A1,A2 G2,G2
50 F1 H3,H4 A1,A1 G2,G2
51 F1 H2,H4 A2,A1 G2,G2
52 F1 H2,H4 A2,A1 G2,G2

Figure T-3 Pedigree screen for the Thomas family. Note that maternally derived alleles are listed first and paternally derived alleles are listed second. Parentheses indicate cases in which maternal or paternal inheritance cannot be determined.



Activity 3
Genetic Anticipation

Focus

Activity 3 introduces students to the phenomenon of genetic anticipation—first, as a newly understood genetic mechanism of interest in its own right, and second, as a way to consider some of the potential problems associated with the rapid increase in our knowledge about the human genome. The activity also illustrates the role that genomic databases play in the dissemination of new scientific findings and identifies some of the serious questions that individuals and society are facing with respect to the interpretation and use of genetic information.

HGP Context

The effort to map and sequence the human genome will not give us all of the answers that we seek to questions about human evolution, human variation, genetic disease, development, behavior, or the interactions between genes and the environment that are associated with these phenomena. It will, however, provide us with an important set of research tools. These tools include not just the human DNA sequence, but also the sequences of several other organisms, as well as all of the technologies developed to generate them. These tools will fuel the work of scientists for years to come.

Already, however, HGP-related research is yielding a wealth of answers to a number of questions of immediate interest, and some of the answers are quite surprising. For example, scientists recently have reported the discovery of a new disease mechanism that involves the expansion of short stretches of DNA to many times their normal length. This new mechanism appears to be involved in at least six human disorders, including fragile X syndrome; a closely related mutation resulting in a fragile site some 500,000 base pairs from the fragile X locus; myotonic dystrophy; Huntington disease; Kennedy disease (spino-bulbar muscular atrophy); and spinocerebellar ataxia type I. All of these disorders involve cells of the neuromuscular system; each is described briefly in the *Glossary*, pp. 59–64.

In the normal alleles of the genes implicated in these conditions, short stretches of DNA, three nucleotides in length, are repeated anywhere from five to as many as 50 times. Individuals who carry alleles that have a normal range of these trinucleotide repeats almost always transmit those alleles stably to their children (that is, there is no change in repeat number from parent to child).

Mutated forms of these genes, however, may contain many more than the normal number of repeats. Although people who carry these mutated alleles still may be healthy, they may not transmit them stably to their children. Instead, the number of repeats may increase. Children of individuals who carry mutated forms may then carry alleles that contain between hundreds and thousands of repeats and these children likely will exhibit symptoms of the disorder. This increase in the number of repeats explains the phenomenon of *genetic anticipation,* a phenomenon in which the severity of certain disorders may increase or the age of onset may decrease with each succeeding generation.

Although the discovery of trinucleotide repeats is fairly recent, the observation of genetic anticipation is not. For example, in the 1950s, the British geneticist Lionel Penrose observed that myotonic dystrophy increased in severity in subsequent generations. He attributed this observation, however, to bias in detection of the disorder. That is, physicians who knew the disorder was present in a given family would look for it more rigorously in subsequent births. Penrose dismissed any biological explanation for anticipation bias, but he did not, of course, have at his disposal the tools of modern molecular biology that have allowed today's geneticists to identify a biological explanation for this phenomenon.

A particularly exciting aspect of this discovery is that the new mechanism was found for the first time in humans, not in an experimental organism such as the mouse. In fact, scientists have seen nothing similar to it in any other species that they have studied. This finding supports the argument that the HGP not only will provide new information about inherited diseases, but also will help us to think in new ways about the fundamental processes of genetics.

Some individuals associated with the HGP, however, are concerned that some of the new information that will emerge from genomic research may have ethical and legal implications that society is not prepared to face. For example, as techniques for identifying and tracking specific DNA sequences are refined, and as these techniques begin to have medical applications, health-care workers will have to deal with complicated scientific, technological, and clinical questions that heretofore may have been of only tangential interest or concern. Likewise, each of us, as "owner" of a genetic heritage, may face serious personal questions about our genetic background and its implications for health, life style, marriage, employment, and reproduction. How will the American public deal with such issues, especially against the backdrop of an incomplete and rapidly changing understanding of the underlying science?

Recognizing these concerns, the HGP has pledged a minimum of three percent of its budget to support research, discussion, and education about the societal implications of the HGP. A diverse working group supported by the Department of Energy and the National Institutes of Health is helping to coordinate these efforts. This group usually is identified by its acronym—ELSI—which stands for Ethical, Legal, and Social Implications related to mapping and sequencing the human genome. The ELSI portion of the HGP supported development of this module as one of its efforts to educate the public about genome research.

Activity 3 introduces students to some of the ethical issues raised by the HGP. The specific phenomenon at issue is genetic anticipation. As the students investigate this phenomenon, they will deal with such problems as the difficulty of interpreting new genetic data properly, the dangers of self-counseling, and an individual's right to privacy with respect to his or her genetic and / or medical profile.

Major Concepts

- □ HGP-related research already is generating data that are of immediate interest and usefulness. An example is the recently reported discovery of the expansion (increase in number) of the trinucleotide repeats that occur in some genes. This is an example of a new disease-causing mechanism that was observed for the first time in the course of studying the human genome. The expansion of trinucleotide repeats explains the long-known phenomenon of genetic anticipation, in which the severity of a genetic disorder may increase or the age of onset may decrease with each succeeding generation.
- □ The HGP raises a number of important questions that will spur research on ethical and public-policy issues. These issues are not unique

to the HGP, but the rapid development of medical and commercial applications of knowledge gained through the HGP may make such questions particularly pressing.

□ Our incomplete and rapidly changing understanding of the human genome makes evaluating ethical and public-policy questions particularly difficult. Information that may be relatively innocuous today may be open to revised interpretation in the light of new findings tomorrow.

□ Genetic information has both individual and family implications. Because it can be used to describe individuals uniquely, we may choose to view it as private, personal information. Because genetic information is shared with other family members, however, it has dimensions that extend beyond the individual. This affects the decisions that we make about its collection, dissemination, and use.

□ Our growing knowledge of genetics raises our sense of connections and obligations to past and future generations and brings with it increased personal and collective responsibility. Acquisition of genetic data about an individual often has implications for other family members. This may challenge traditional notions of privacy and confidentiality.

Student Objectives

As students complete this activity, they should

□ understand the phenomenon of genetic anticipation as an interesting variation on traditional Mendelian inheritance;

□ understand the amplification of trinucleotide repeats as a new disease mechanism that was discovered in humans rather than in a model organism;

□ search a registry database for information about the genetic profile and medical history of a young woman and her extended family;

□ search a research database for information about the fragile X gene and its associated phenotypes and mode of inheritance;

□ understand that an individual's genetic data often have implications for other family members;

□ recognize some of the difficulties of interpreting personal genomic data in this situation; and

□ identify some types of misuses and abuses of personal genomic information that would be possible if these data were made freely available to anyone interested.

Science Process Skills

□ Observing

□ Comparing

□ Communicating (orally and in writing)

□ Using the computer as a research tool

Materials for a Class of 30

□ 30 sets of student text, pp. S-25–S-26. Because students will complete their written work on their own paper or on the worksheets provided, students can reuse these pages.

□ 30 copies of Worksheet 9, *Genetic Anticipation and Trinucleotide Repeats.*

Advance Preparation

□ Set the software on each computer to *Genetic Anticipation.* You will find instructions for how to do this on p. 56. The password is <McClintock>. You may wish to have your students perform this process; we provide a blackline master listing the steps required at the end of Activity 1 (BLM T-3).

Estimated Time

One 45-minute class period

Introduction

Activity 3 is designed to accomplish a number of objectives, some related to the science and technology of the HGP and others to the ethical questions that the science and technology raise. An underlying pedagogical goal of the activity is to help students see how the science, the informatics, and the ethical issues related to the HGP intersect, and to understand that this intersection makes public-policy questions related to the HGP both immediate and urgent.

The first objective of the activity is to give the students an opportunity to explore genetic anticipation as a phenomenon that extends the classical Mendelian assumptions about transmission and expression. The intent here is to communicate to students the excitement of research on the human genome, to give them a sense of some of the HGP's concrete outcomes, and to help them understand that there is much more to learn about human genetics than we have discovered already.

A second objective of the activity is to illustrate for students how quickly our understanding of specific biologic phenomena can change and to help students see that new discoveries do not necessarily replace or invalidate old understandings. Rather, such discoveries add to the incredibly complex biologic picture already in place. Thus, our new understanding of the expansion of trinucleotide repeats as an explanation for genetic anticipation does not lead scientists to discount the importance of Mendelian inheritance. Instead, it reminds us again that the mechanisms of inheritance are richer and more complex than a picture that is based *solely* on single genes operating in strict Mendelian fashion would suggest.

The activity also is designed to help students understand the role of electronic databases in the rapid spread of new information about the human genome. Students should understand that as the volume and the complexity of the information about the human genome increase, scientists will depend more heavily on electronic databases, not only for information storage but also for its rapid dissemination of research findings to the rest of the scientific community. Remind students that the major research databases associated with the genome project are updated *daily* and that a successful scientist must stay abreast of new developments. Point out, as well, that new research results sometimes have unexpected implications, both for science and for society. Although it is rare that new data are as dramatic as the evidence for a new genetic phenomenon, the regular and timely exchange of information among involved laboratories and individuals is an important aspect of genomic research.

Finally, the activity completes the process of raising the students' sensitivities to the importance of genetic data as *family* data and to issues of privacy and confidentiality as they relate to genomic information. In this sense, the exercise acts as an important bridge from the earlier activities, which focus on the science and the informatics of the HGP, to Activities 4 and 5, which deal with the ethical and public-policy issues that the science and informatics raise. In particular, Activity 3 will help students recognize that genetic data sometimes have significant implications for an individual and for a family and that these implications can change as new findings emerge. Privacy is not the focus of Activity 3; nevertheless, at its conclusion, students should understand that genetic information can change lives and relationships. This will help create a context within which, in Activities 4 and 5, students will consider a variety of ethical and public-policy issues related to the release of such information from genetic registries.

Related Information in the Teacher Narrative

You will find additional background information in the following pages of the teacher narrative:

- □ expansion of areas of trinucleotide repeats as a recent discovery associated with the HGP (p. 3)
- □ genetic information as family data (pp. 36–40)
- □ issues of privacy and confidentiality as they relate to genetic information (pp. 36–37, pp. 43–44)

Suggestions from the Field Test

- □ To allow more time for in-class discussion, you may wish to ask the students to read Worksheet 9, *Genetic Anticipation and Trinucleotide Repeats,* as homework.
- □ This activity offers an excellent opportunity to introduce students to the role of genetic counselors and medical geneticists in the health-care system. You might invite a genetic counselor into your class to talk with the students after they complete this activity. Alternatively, you might ask such a visitor to work through the exercise and the discussion questions with you and your students.

111

Annotated Student Activity

Activity 3 is structured as an open-ended exercise in which the students work in groups to investigate a variety of issues related to Joy Major's dilemma (see below), and then reconvene as a class to consider a final set of summary questions.

Because the phenomenon of genetic anticipation and its relationship to trinucleotide repeats may be new to some teachers, we provide extensive annotations to both the student questions below and to the fictitious newspaper clipping.

PROCEDURE

Remind your students that they have been imagining that they live in a small town some time in the future. Remind them as well that the increase in information available about the human genome has had a significant impact on this town, as it has on many other communities.

Imagine that it is now six months after the local health authorities first created the LGD and allowed public searching of a small subset of its data. Joy Major is 21 years old, is working her way through the local community college, and, until yesterday, was thinking about becoming a high school mathematics teacher. She finds school challenging and has to work hard to maintain her grades, but she thinks that this will help her be a good teacher. She has promised herself that *she* will never forget how it feels to be confused.

Two weeks ago, Joy was assigned a genetics project in her biology class. This project involved drawing a pedigree for her immediate and extended family and tracing the inheritance patterns of several genetic traits through this pedigree. Joy had been enjoying this exercise until conversations with her older sister, Leah, revealed that doctors had told Leah that her son's mental retardation likely was inherited. When Joy expressed some surprise at this news, Leah told Joy that when her son was diagnosed, the tests to trace the inheritance of the condition were hard to interpret. Because of this, Leah had declined further testing for herself and her son and had not told the rest of her family for fear of worrying them unnecessarily.

Emphasize that it is now six months after the LGD first was created.

Before the discovery of the expansion of trinucleotide repeats at the fragile X locus, counseling of fragile X families was extremely difficult. This was primarily because of the reduced penetrance[1] and variable expressivity[2] of the trait and because genetic analysis of the disorder was based on cytological tests designed to identify the presence of a fragile site (a gap or defect observed in the continuity of a stained chromosome) in band 27.3 of the long arm of the X chromosome. Because only about 50 percent of all known carriers of fragile X exhibit the fragile site, however, it was hard to say which women in an affected family were at risk for producing affected children. In addition, although male fetuses with a fragile site usually would be retarded, there appeared to be no way to predict whether a female fetus with a fragile site would be retarded or unaffected.

Even now, although health-care workers can determine the exact number of CGG repeats an individual carries, forecasts about the appearance of the related phenotype are not completely clear. This may reflect the capacity that the areas of trinucleotide repeat have for change. Not only can the number of repeats change from parent to child, but the number of repeats also can change during mitotic division. This mitotic instability means that individuals with a fragile X mutation can be somatic mosaics for the mutation (that is, they may show different numbers of repeats in different tissues). If the effect of the protein product that is associated with fragile X is largely restricted to the cell in which it is produced, the numbers of repeats that occur in particular cells of the body (for example, the neurons) may determine the phenotypic expression of the mutation in single individuals. Thus, the dynamic nature of the mutation within single individuals and the phenomenon of X-chromosome inactivation in females may explain the incomplete penetrance and variable expression found in this disorder.

After Joy talked with Leah, Joy became curious about what scientists might have discovered about the inheritance of fragile X since her nephew was born and she decided to do some research on her own. First, she found the newspaper clipping that your teacher has given you. When she read the article, she became quite upset.

Distribute one copy of Worksheet 9, *Genetic Anticipation and Trinucleotide Repeats,* to each student. Give the students about ten minutes to read the article and then discuss the questions in Step 1 with the class. Alternatively, you may wish to allow the students to discuss the questions in Step 1 in small groups before you begin the class discussion.

You will find an extended discussion of the phenomena described in the newspaper article at the end of the annotated student activity (pp. 106–107).

1. Do you see anything in the article that suggests why Joy might be so upset?

HINT: Make sure that you understand what the article says. Answering each of the questions below may help you with this.

a. With what inherited condition is the fragile X gene associated?

The fragile X gene is associated with inherited mental retardation.

b. What is unusual about the manner in which this condition is inherited?

The trait is inherited as an X-linked dominant, but the associated phenotypes do not always appear when the altered form of the gene is present. Likewise, the associated phenotypes can increase in severity from one generation to the next, as in the appearance of mental impairment in the sons and daughters of a woman who receives a mutated gene from her father, but seems to be otherwise unaffected by it.

c. What is unusual about the DNA sequence of the fragile X gene, in comparison with many other genes?

The fragile X gene contains an area where a three-base sequence appears over and over again without any intervening bases. The repeated three-base sequence found in the fragile X gene is the trinucleotide CGG.

d. How many trinucleotide repeats normally occur in the fragile X gene?

Normally, individuals carry genes that have between six and about 50 copies of this trinucleotide.

e. What is different about the number of these repeats in individuals said to carry a "premutation" for fragile X?

Individuals with a fragile X premutation may show between about 50 and 200 copies of the repeat.

f. How can the number of repeats change as premutations are transmitted from one generation to the next? What are the possible phenotypic consequences of such changes?

The number of repeats can increase as premutations are transmitted from one generation to the next. People with full fragile X mutations may show as many as 1,000 CGG repeats. Mental retardation that ranges from mild to severe can accompany both the premutated form of the gene and the full mutation.

If you already have discussed Huntington disease (HD) with your students, point out that it, too, shows genetic anticipation (that is, the symptoms can get more severe with successive generations and the age of onset can decline). The underlying mechanism in HD also appears to be the expansion of areas of trinucleotide repeat. In HD, the trinucleotide involved is CAG.

2. Follow your teacher's instructions for breaking into small work groups around the computers to discuss Joy's situation more closely. Use Questions 3-7 below to guide your discussion. Consult the databases as required and be prepared to join the rest of the class in an open discussion by the time your teacher specifies.

As stated earlier, this activity is structured as an open-ended exercise. For this reason, we do not provide the students with specific directions about how to use the LGD and the NGD databases. The objective here is to have students use these databases as Joy Major does and discover the information as she does. Guide them in this direction.

3. When Joy first read the clipping, she remembered that fragile X was one of the genes that was recorded in her LGD file. As soon as she got the chance, she looked up her file and then cross-referenced it to the information in the NGD.

a. What is Joy's genotype for fragile X?

Joy carries the genotype F1/F3 for fragile X.

b. From whom did she inherit the allele that concerns her?

She inherited her F3 allele from her father. (The NGD describes the F1 allele as containing a normal number of repeats. In contrast, the F3 allele contains a much higher number of repeats.)

c. Why may this information have frightened her?

The NGD identifies this allele as carrying repeats at the high normal or low premutation level. Joy probably is worried that her father carried a premutation for fragile X and that she is at risk for bearing mentally retarded children. She also may be questioning her own mental capacity because female carriers sometimes show some learning difficulties.

4. After Joy examines her own file, she looks up her siblings' genotypes for the fragile X gene.

a. What genotype with respect to fragile X does Leah carry?

Leah has the same genotype for fragile X that Joy carries: F1/F3.

b. How does this compare with Joy's genotype? What might this mean for Joy?

Leah and Joy have the same genotype for fragile X. Like Leah, Joy could have a child with fragile X syndrome.

c. Consider Joy's other sister, Anna. What is her situation with respect to fragile X?

Anna has the same genotype for fragile X that Joy and Leah carry: F1/F3. Each of the females in this family is in the same genetic situation.

d. What is their brother's genetic status with respect to this gene?

Joy's brother, Joe, inherited an F1 allele from his mother, so he is not at risk. Your students may decide to examine the records of Joy's sisters more closely to determine whether there is any hint of mental deficiency in Joy's family other than Leah's son. Your students will find that Leah has a degree in chemistry. Anna, in contrast, may not have been a good student. Point out to your students, however, that this is *not* evidence of mental impairment or of fragile X syndrome.

5. Joy also searches the LGD for evidence of the fragile X syndrome in other families in the community.

a. What does she find?

HINT: You may want to consider searching the LGD by gene symbol to retrieve this information.

If your students are not sure how to conduct this search, suggest that they first search the LGD for the F3 allele (select *Gene Symbol* in the *TYPE* box and *F3* from the *VALUE* list). This will retrieve a list of all of the people in the LGD who carry one or more F3 alleles. If the students click on a name on the list, the computer will display the record for that person. Remind the students that they can return to the list by clicking on the button that reads *RETURN TO LIST*. They can search the LGD on the F4 allele by following the same general steps. (**Windows Users:** Double click on a person's name on the list to retrieve his or her file. Click on *BEGIN SEARCH* to return to the list.)

The LGD shows that one other family in the town also carries some F3 and F4 alleles for fragile X. Accessing the personal records of individuals carrying these alleles reveals some cases of learning difficulties.

b. How might this information influence Joy's attitude toward these other community

members? How might this information affect Joy's feelings about herself?

Joy may be starting to feel that she is not "normal," and this may be affecting her self-esteem.

6. One of Joy's responses to the information in the LGD and the NGD is to wish that she and her family had not agreed to allow their personal files to be made public, *even* for educational purposes. Do you think that she has a point? Why or why not?

Answers will vary. Remind your students that when the LGD was placed online, scientists knew very little about the significance of the allelic sequences associated with fragile X. Now, however, six months later, the NGD reflects the new knowledge that scientists have generated as a result of continued research. According to the scenario we have established to explain the LGD, each person gave his or her consent to have personal genetic data entered into the database. Perhaps Joy should have realized that as our understanding of the human genome changes, the implications of an individual's genetic profile also may change. This might have led her to refuse her consent, even though when she was asked, she was not concerned about her particular profile.

If your students have difficulty with this concept, suggest that they think back to Activity 2 and describe what the NGD said then about the phenotypic consequences of the various fragile X alleles. (The student in each group who chose fragile X to study should remember that the NGD said "No information is available about the phenotypic significance of this allele.") Ask your students what has happened in the intervening six months to change the information available in the NGD. Then ask them what decision Joy might make now about giving her consent for her genetic profile to be made public.

7. Examine Joy's genotype with respect to fragile X very closely. What does it tell you? What does it *not* tell you? Based on the data that you have available, what can you say about the probability that Joy carries a fragile X premutation?

This is a key question. Joy's genotype simply tells us that Joy has either a high normal or a low premutation number of repeats. It does not tell us in which category Joy falls. Based on the data available in her file, we cannot say for sure what Joy's genetic situation is. Leah's report that her son suffers from fragile-X-related mental retardation, however, suggests that Joy's father carries a premutation and that Joy's children would be at risk.

8. Follow your teacher's instructions for discussing the following questions as a class.

Use the following questions to guide discussion, but not to restrict it. Encourage the students to answer specifically, citing information from the clipping and from the database as appropriate.

Questions for Discussion

1. What is the dilemma that faces Joy?

Joy is trying to plan her future in the face of unsettling, but incomplete information. The situation is emotionally charged and value laden.

2. What is unusual about the fragile X gene that makes Joy's situation particularly confusing?

The phenomenon of the expansion of trinucleotide repeats and the variable penetrance and expression of the mutation makes precise predictions about the implications of her genotype difficult.

3. What role did the rapid increase in scientific understanding play in Joy's dilemma?

Before the discovery of the relationship between genetic anticipation and trinucleotide repeats, Joy's sequence data were unremarkable. The discovery of this relationship and its rapid dissemination through genomic databases changed this situation very quickly.

4. What role did electronic databases play in Joy's dilemma?

The availability of electronic databases made the connection between Joy's sequence data and the clinical implications easy to determine. In addition, the data in this case are accessible to the public.

5. Joy attempted to interpret complex genetic information without qualified help. What role did this play in her dilemma? Why is it difficult to define "normal" in this situation?

Joy was drawing conclusions that might not be valid. It would help her to seek professional advice as she attempts to interpret these data. "Normal" is always a difficult notion to define, especially with respect to mental capacity. The boundary, for example, between "low normal" and "mentally retarded" is very difficult to define clinically. Furthermore, tests for mental capacity are open to wide interpretation and many have been criticized for inherent biases or for lack of validity. Remind your students as well that the fragile X gene is only one of many factors affecting intelligence.

Likewise, the uncertainty related to the probability that Joy would bear a child with a full fragile-X mutation likely is unsettling to Joy. Joy should seek help from qualified professionals to determine her risk for bearing children with fragile X syndrome.

6. What implications does Leah's son's mental retardation have for her sisters? In what sense are genetic data private and in what sense do we share such data with others? Explain.

Leah's genetic data are hers, but they also have implications for her sisters, who would bear the same risk. In other words, genetic information is, in important ways, family information. In addition, this information can have profound implications for generations beyond Joy's children and those of her sisters. Genetic counselors often face the problem of encouraging patients to share unsettling genetic information with other family members. Sometimes, genetic explanations for a family problem challenge mistaken, but longstanding explanations that have become an accepted part of the family's culture.

7. If Joy consulted a genetic counselor, she would learn, among other things, that there are prenatal (before birth) tests to detect the presence of the fragile X mutation in the developing fetus. How might this information affect Joy as she thinks about her future?

The responses will vary. Certainly it is a good idea for Joy to seek professional advice. Prenatal diagnosis can provide additional information to help people make informed reproductive decisions, but the information itself raises difficult questions. Prenatal diagnosis, of course, is entirely optional—a fact the counselor would point out to Joy. The genetic counselor also would review all of the options related to a diagnosis of potential mental retardation: carry the pregnancy to term and raise the child oneself with appropriate educational programs and support; consider institutionalization, if, for example, the retardation is severe; consider offering the child for adoption; consider therapeutic abortion.

All of these options are complicated by the incomplete nature of our knowledge about the fragile X mutation, by the variable expression of the phenotype, and, of course, by the personal values of the people involved—in this case, Joy, her family, and a future spouse, if she decides to marry. Genetic counseling programs are designed to help individuals, couples, and families with these difficult issues and decisions. Although the counseling team is supportive and provides as much information as possible, the team members do not make the decision for the patient. That responsibility ultimately must reside with the people who will live with the consequences of the decision.

FINAL NOTE: To conclude this activity, ask your students what information this exercise has given them that may help them better understand the genetic incompatibility involving Walter Raynes (Activity 2). If they reexamine Walter's genotype in relation to those of his parents and grandparents, students could recognize that genetic anticipation is the missing explanation for the apparent discrepancy in this family. Walter inherited his F4 allele from his mother, who carried an F3 premutation. Walter's sister, in contrast, inherited their mother's *other* X chromosome. Point out to your students that data that look confusing now may be explained in the future as scientists make new discoveries. Ongoing genetic research almost will certainly raise many serious and difficult issues. It also, however, will *answer* many important and difficult questions.

ANNOTATIONS TO WORKSHEET 9
Genetic Anticipation and Trinucleotide Repeats

Paragraph One: The newspaper article focuses on the changes that can occur in the DNA of the fragile X gene from one generation to the next. Make sure that students understand that the potential always exists for the DNA to change during the normal processes of reproduction. In fact, these changes, passed on from their occurrence to subsequent generations, are central to evolution. The expansion of areas of trinucleotide repeats then, does not constitute a new mechanism because it has to do with changes in the DNA from parent to child. Rather, this phenomenon is novel in the *degree* of that change (as you will see below, the change can involve thousands of bases), and in the *relative predictability* of that change (though these changes do not always occur, they nevertheless occur with a much higher frequency than might be expected in a randomly mutating gene).

Paragraphs Two and Three: The core of science lies in seeking explanations for our observations about the natural world. You may wish to illustrate this concept for your students by tracing some of the history of the research on fragile X syndrome. More than 60 years ago, researchers noted that males outnumber females by about 25 percent among the institutionalized cases of mental retardation. In the 1940s, researchers suggested that this observation could be related to an X-linked inheritance pattern, and, in 1969, a cytological defect was found on the long arm of the X chromosome of four retarded males in one family. Because cytological defects of this type were known to be associated with chromosome breakage, this defect was dubbed a *fragile site* and the related X-linked mental retardation was named *fragile X syndrome*. (Approximately 20 other heritable fragile sites are known to exist on other chromosomes, but none of these has been associated with a particular phenotype.) It was not until the 1970s that scientists recognized fragile X syndrome as the most common form of inherited mental retardation.

Data on the unusual inheritance pattern displayed by the fragile X gene accumulated slowly during the years after its identification. It was not until 1991, however, that reports of an area of CGG repeats within the fragile site appeared in the literature and that researchers were able to demonstrate the direct involvement of these repeats in the expansion of this area that scientists saw in fragile X patients. The discovery of this new disease-causing mechanism suggested that similar mechanisms might be operating in other disorders that show genetic anticipation, and early in 1992, the genetic lesion in myotonic dystrophy was identified as an expansion of an area of CTG repeats. This discovery was followed in 1993 by the demonstration that trinucleotide mutations also are involved in Huntington disease, spinocerebellar ataxia, Kennedy disease, and in the mental retardation associated with a mutation that is located distal to the fragile X site on the X chromosome.

In classical terms, the fragile X gene has been described as an X-linked dominant with variable expression and reduced penetrance. Early observations revealed that the degree of mental impairment among individuals displaying the fragile X phenotype can show significant variation. Likewise, children who have inherited the same at-risk genotype from their mother may be affected or unaffected. In the case of female children, this observation may be explained by X chromosome inactivation (for example, if it is the X chromosome bearing the fragile site that is inactivated in the tissues that express the fragile X phenotype, the female may be unaffected). In the case of males, the reduced penetrance is harder to explain, but it may result in part from the mitotic instability described in the annotation to the introduction to the student procedure (p. 101).

Figure 6: Be sure to tell your students that the general scheme shown in the diagram is *not* an actual pedigree and does not represent accurately the proportion of individuals at risk for certain genotypes or phenotypes. Emphasize that it only illustrates the types of combinations that can result from one scenario involving fragile X (the case in

which a male of normal intelligence passes the premutation to a daughter). Point out the variability involved: the woman who receives the premutation from her father may bear sons and daughters who carry the gene and who also show mental impairment, sons and daughters who carry the gene but do not show mental impairment, and, as a result of passing on her *other* X chromosome, sons and daughters who do not carry the fragile X gene at all. Likewise, among those children who *do* show the associated phenotype, mental impairment can range from moderate to severe. This type of variability makes the challenge of counseling a fragile-X-bearing woman particularly complex.

As you review this information with your students, be certain to distinguish between

□ genetic anticipation—the long-standing *observation* that some genetic disorders may increase in severity or decrease in age of onset with each generation, and

□ trinucleotide repeats—a recently discovered *genetic mechanism* that helps to explain the observation of genetic anticipation.

Paragraphs Four, Five, and Six: Your students should understand that the probability of a child being affected with fragile X depends upon the number of repeats carried by the mother and upon whether that number increases as the gene is passed to her child. If the mother has inherited a premutation (a gene with a larger-than-normal number of repeats) from her father, this gene can mutate further as it passes through her, and the number of repeats will tend to increase with each succeeding generation thereafter. Scientists do not yet understand why an inherited premutation does not increase in length as it is passed from a father to a daughter, but only from that daughter to *her* children.

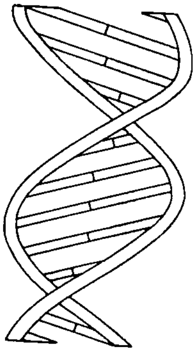
Recent research[3] suggests that in addition to the varying number of CGG repeats, the degree to which the DNA in this region is methylated also varies. ("Methylation" refers to the process by which particular cytosines in a DNA molecule are enzymatically converted to 5-methyl cytosine by the addition of a methyl group [$-CH_3$] to carbon 5 in cytosine. Methylation is explained in more detail in the *Glossary*.) Several investigators have reported that unaffected transmitting males and their daughters (that is, individuals carrying fragile X premutations) have unmethylated DNA in this area of their active X chromosomes. The full fragile X mutation apparently occurs when the number of trinucleotide repeats expands to more than 200 with accompanying methylation. Investigations into the protein product encoded by the fragile X gene have revealed that the CGG repeats lie in an area of the gene that is transcribed (the associated mRNA molecule is of different lengths in individuals with different repeat numbers), but is not translated into protein (no differences in lengths of the protein product are observed in such individuals). Expansion of the number of CGG repeats beyond approximately 230 usually is accompanied by methylation of the DNA at the CGG-containing end of the gene. This methylation may eliminate expression of the gene—methylation of the fragile X gene correlates well with loss of transcription, and absence of the protein product correlates well with the appearance of the associated phenotype—but a direct causal relationship has not yet been established.

NOTES:

[1] The *penetrance* of a trait is defined as the proportion of individuals having a specific genotype who actually express the related phenotype. A trait that shows reduced penetrance, then, is not always expressed in individuals known to carry the associated genotype.

[2] The *expressivity* of a trait is defined as the degree to which a penetrant gene is expressed phenotypically. A trait that shows variable expressivity may manifest itself differently in different individuals (individuals in whom the gene is penetrant) and even may manifest itself differently in different parts of the same individual. The range of expression of some traits extends continuously from full expression to almost no expression of phenotypic characteristics.

[3] McConkie-Rosell, A., et al. (1993). Evidence that methylation of the FMR-1 locus is responsible for variable phenotypic expression of the fragile X syndrome. *American Journal of Human Genetics* 53:800-809.



Activity 4 Who Should Control Information about My Genes?

Focus

In this activity, students use the skills of ethical reasoning to analyze issues related to access to and use of genomic data. Students consider several possible levels of disclosure of the results of the DNA testing that was initiated in Activity 1 and determine whether the new data should be added to the LGD. Activity 4 sets the stage for Activity 5, where students consider proposed public policies about access to and use of genomic data.

HGP Context

By the time the students reach Activity 4, they will be familiar with research and registry databases, and they will have encountered central ethical concepts, including privacy and confidentiality, that might restrain access to data in a genetic registry. A right to privacy and the obligation of confidentiality are based on respect for autonomy, that is, the capacity of each individual to be self-determining in thought and behavior.

In some cases, however, access to data in a genetic registry may provide beneficial consequences. Put another way, overriding a right to privacy or an obligation to confidentiality may be justified when rights or consequences are in conflict or when exercising a right or obligation would result in serious, irreversible harm to the self or others.

In this activity, the students explore pertinent ethical concepts related to the access to and use of information from genetic registries. Specifically, students are asked to decide the extent to which the results of the genetic testing that they authorized in Activity 1 should be disclosed publicly. Discussions focus on identifying possible responses to a range of disclosure options and require students to consider how to balance ethical commitments tied to individual privacy against commitments related to consequences or outcomes. As the students learned in Activities 1-3, genetic information is family-related information, and decisions by and about one family member may raise challenging ethical dilemmas for each of the members of a particular family.

Major Concepts

- □ *Privacy* is the right to control access to information about oneself. *Confidentiality* is an obligation of those who obtain information about other individuals to protect the privacy of that information by controlling access to it.
- □ Overriding a right to privacy may be justified when privacy rights conflict or when exercising a privacy right would lead to serious harm to the self or others.
- □ Because genetic data are family data, failure to

protect the privacy of genetic information can create problems for the individual and for other members of the family.

□ Databases can create problems of privacy and confidentiality because it often is difficult to protect data in electronic databases.

□ Ethical inquiry involves accumulating information, evaluating information, making arguments, and analyzing arguments.

Student Objectives

As students complete this activity, they should

□ use the skills of gathering information, evaluating information, and making and analyzing arguments as tools for ethical inquiry;

□ evaluate ethical issues related to registry databases;

□ take and explain a position on whether respect for individual privacy or the pursuit of valued outcomes or consequences is the more important ethical consideration regarding the release of genetic information; and

□ appreciate that other students may have well-argued, alternative views concerning these issues and that the response to well-founded disagreement should be thoughtful respect and continued dialogue.

Science Process Skills

□ Gathering data
□ Describing
□ Inferring
□ Synthesizing information and knowledge
□ Communicating (orally and in writing)

Ethics Process Skills

□ Gathering information
□ Evaluating information
□ Making arguments
□ Analyzing arguments
□ Appreciating and responding thoughtfully to disagreement

Materials for a Class of 30

□ 30 sets of student text, pp. S-29–S-30. Because students will complete their written work on their own paper or on the worksheets provided, students can reuse these pages.

□ 30 copies of Worksheet 10, *Genetic Registries, Information, and Privacy*. Assign Worksheet 10 as homework in preparation for Activity 5.

□ 30 envelopes containing 3 cotton swabs each, with signed instructions as to whether to test or not. These are the envelopes that your students prepared in Activity 1 and that you will use to report their individual test results.

□ Overhead transparencies of the results of the DNA testing (BLM T-4), the options for disclosure of the test results (BLM T-5), and the instructions for interpreting the results of the DNA testing (BLM T-6).

□ One 9″ x 12″ envelope in which to place the transparency that displays the fictitious results of the DNA testing (BLM T-4).

□ Butcher block paper, markers, and transparent tape (optional).

Advance Preparation

The following instructions assume that your class completed the simulated DNA testing described in Activity 1, *Genetic Registries* (pp. 75–82).

□ Leave the software on each computer set to *Genetic Anticipation*. (There is no special database setting for Activity 4, although the students will use the computer to gather information about the genetic disorders involved.)

□ Make the necessary student copies and overhead transparencies.

□ Place the overhead transparency that lists the combined test results (BLM T-4) into an envelope and seal it. Label the outside of the envelope with the class name, section number, and the fraction of students who chose to be tested. For example, if 23 of 24 students chose testing, write 23/24 on the outside of the envelope. You will need this number during the introduction to the activity.

□ Complete the following steps to prepare the envelopes containing the individual test results.

1. Color one swab in the envelope(s) signed by Aisha Jacobs *red;* this indicates carrier status (heterozygosity) for cystic fibrosis. The other two swabs for Aisha remain white. Seal the envelope(s) containing the swabs.

NOTE: *If no students were assigned this identity, or if all of the students assigned this identity declined the testing, you may designate either George Jacobs* *or Katherine Schmidt as the one carrier for cystic fibrosis in the sample population and prepare his or her swabs accordingly.*

2. Color one swab in the envelope(s) signed by Kay Raynes *blue;* this indicates carrier status (heterozygosity) for sickle cell disease. The other two swabs for Kay remain white. Seal the envelope(s) containing the swabs.

   NOTE: *If no students were assigned this identity, or if all of the students assigned this identity declined the testing, you may designate either Anna McCarthy or Christina Chen as the one carrier for sickle cell disease in the sample population and prepare her swabs accordingly.*

3. Color one swab in the envelope(s) signed by Bob Thomas *green;* this indicates the allele for familial hypertrophic cardiomyopathy. The other two swabs for Bob remain white. Seal the envelope(s) containing the swabs.

   NOTE: *If no students were assigned this identity, or if all of the students assigned this identity declined the testing, you may designate either Cal Thomas or Mary Jo Wray as the one individual with familial hypertrophic cardiomyopathy in the sample population and prepare his or her swabs accordingly.*

4. Leave all of the swabs for all other individuals in the sample population *white.* The assumption is that any swab that remains white shows a negative result (absence of the abnormal form of the gene in question). As indicated above, this is the case for all but *three* of the fictitious individuals in the test population. Seal these envelopes.

SPECIAL INSTRUCTIONS: Depending on how you assigned fictitious names to your students in Activity 1, it is likely that several students in your class will have the same fictitious identity. Because the decision to be tested is voluntary, students who share the same fictitious identity may respond differently to the invitation to be tested. If this happens, explain to your students that this is an unavoidable result of the manner in which Activities 1 and 4 are structured and suggest that the class deal with it in one of the two following ways:

a. decide that each student will behave autonomously with respect to Activity 4 and, to preserve the full experience of the activity for each student, promise not to share test results with each other; *or*

b. decide that all the students in the class who share one identity will make collective decisions on behalf of this person. This will require that all of these students reach consensus on each key step before going on to the next. This option may slow the activity if students have difficulty reaching consensus, but it may provoke some interesting dialogue.

Estimated Time

One 45 minute class period

Related Material in the Teacher Narrative

You will find additional background information on the following pages of the teacher narrative:

□ ethical issues related to registry databases (pp. 36–40)

□ legal protections of privacy of genetic information in the United States (pp. 43–44)

□ dealing with values and controversial issues (pp. 44–45)

Suggestions from the Field Test

□ Ask your students to investigate whether privacy legislation concerning genetic information or computerized medical information is in force or is pending in their state. Recommend that students call their elected state officials, state medical society, or state department of health to locate information about such legislation.

□ Suggest that interested students interview their school principal about databases that may be used in their school or school system and about the protections that are in place to discourage inappropriate use of information about students.

Annotated Student Activity

At the end of Activity 1, your teacher asked you to decide whether to authorize further testing of your fictitious person's DNA. The purpose of this testing was to determine his or her genetic status with respect to the genes for cystic fibrosis (CF), sickle cell disease (SCD), and familial hypertrophic cardiomyopathy (CM).

In this activity, you will consider several options for how to handle the test results.

This activity is organized into two conceptual parts. In the first part, students consider whether and how to disclose the test results (first, the anonymous, combined data and then the individual test results). This portion of the activity asks students to explore arguments for and against maintaining the confidentiality of the results of genetic testing and raises a number of important questions about personal and family privacy and well-being.

In the second part of the activity, students consider whether the individual test results should be entered into the LGD. In this portion of the activity, students examine some of the special ethical issues that arise from the specific use of electronic databases as mechanisms for the storage and dissemination of personal data.

Procedure

Begin by saying something such as the following:

"At the end of Activity 1, 00/00 [*insert the figure written on the outside of the envelope*] of you chose to have your fictitious person tested for cystic fibrosis, sickle cell disease, and cardiomyopathy. We now have received the results from GeneTest, Inc., a commercial firm that does DNA testing. Your individual results are sealed in the envelopes you prepared. The group results are sealed in this separate envelope. In this activity, we will decide how to handle this information. Before we begin to discuss our options, however, take a few minutes to find out what the NGD says about the alleles involved."

1. Good ethical analysis of science-related issues requires a solid, accurate understanding of the relevant science. Follow your teacher's instructions for forming work groups around the available computers. Consult the NGD for descriptions of the disorders listed above and answer the following questions for each.

Give your students 10 minutes to locate and discuss the required information. *Note that there is no special database setting for Activity 4; any of the settings used for Activities 1-3 will make the required information available to students.*

a. What are the medical consequences (if any) for a person who discovers that he or she carries one copy of this allele?

There are no immediate personal medical implications for an individual carrying one allele for cystic fibrosis or one allele for sickle cell disease.

In contrast, discovering that one has the allele for familial hypertrophic cardiomyopathy carries with it significant immediate implications. These implications include risk for sudden heart attack and the accompanying adjustments in diet and activity levels that might be required to reduce that risk. Familial hypertrophic cardiomyopathy is an important cause of sudden, nontraumatic death in the young. The molecular basis in some patients is a mutation in one of the cardiac myosin heavy-chain genes. Some patients carrying the mutation are asymptomatic; others may die in infancy. Sudden death in known carriers occurs at a rate of about 4 percent per year and cannot yet be predicted by any known phenotypic marker. Even among patients carrying the same mutation, the severity and phenotypic expression of the disease vary. This observation suggests that other genetic or environmental factors also are involved in determining a patient's actual phenotype.

b. What are the potential reproductive implications (if any) for a person who discovers that he or she carries one copy of this allele?

The reproductive implications for an individual carrying one allele for any of these three traits have to do with the risk of passing the allele on to his or her offspring. This risk for each trait is 50 percent for each offspring.

Because cystic fibrosis and sickle cell disease are expressed as recessive traits, children who receive only one such allele from their parents would be expected to bear no medical consequences. Should the parent's spouse, however, also be a carrier for the same recessive allele, the risks associated with cystic fibrosis and sickle cell disease change. In this case, each potential offspring would bear a 25 percent risk of inheriting two copies of the harmful allele. In this event, the child would be affected by the disorder in question, although the degree of severity of its expression would be difficult to predict.

Because hypertrophic familial cardiomyopathy is a dominant trait, the reproductive implications for a person who carries one allele for the condition amount to a 50 percent chance of having a child with the same potential for sudden heart attack as the parent. Some students may argue that it is not likely that any of them would unknowingly carry this allele because it probably would have been detected or expressed in the parent from whom they inherited the allele. Point out that because the allele shows variable expression, it is possible that one parent *does* carry the allele, but does not show any related symptoms. The gene also appears to have a relatively high rate of spontaneous mutation, so it is possible that an individual might carry an abnormal allele even though neither parent did.

2. Follow your teacher's instructions for participating in a class discussion about the advantages and disadvantages of handling the test results in several different ways. First, you will explore the following three options:

- **□ report only the anonymous, combined data (Option 1)**
- **□ report each set of individual results only to the person tested (Option 2)**
- **□ announce the individual results openly or publicly to all (Option 3)**

Your class will vote after the discussions about Options 1 and 3.

You should guide the discussions for this part of the activity. Note that the exercise is organized to encourage students to eventually consider all four options and that the options have been sequenced so that they move toward increasingly public disclosure of the individual test results. Initially, the students will explore three options and then in Step three will consider a fouth option. The intent here is to build a structure within which students can give appropriate consideration to a range of possibilities from no more than anonymous disclosure of the combined results (Option 1), to formal, public disclosure of the individual results (Option 3). The overhead transparencies titled *Options for Disclosure of DNA Test Results* (BLM T-5) will help you organize the discussion.

Disclosure Option 1
Report only the anonymous, combined data.

Introduce the discussion of Option 1 by saying something such as the following:

"Now let's consider several options for how to handle the test results. First, I can give you the total number of people in this group who have been identified as having the alleles for CF, SCD, or CM. There will be *no names* attached to these data. How do you feel about my making this information public? What are some reasons to make this information public? What are some reasons to protect this information?"

List the reasons on the transparency, the chalkboard, or on butcher block paper as students provide them. Figure T-4 provides some reasons *to make these data public* and some reasons *not to make these data public.*

Allow the students to decide by majority vote whether to disclose the combined data. Implement their decision by discarding the envelope or by opening the envelope and displaying the overhead that is inside. *(The data from GeneTest, Inc. are displayed on BLM T-4, which you should have prepared earlier and sealed into the envelope.)*

If the students decide to have the data disclosed, allow the students time to think about the data

Make the Combined Results Public

The combined data may help scientists and health-care specialists understand the degree to which this population is at risk for certain genetic disorders. This may help them plan more effectively for genetic services in the community.

These data might add to the base of information scientists have available for conducting research into these conditions.

Genetic information belongs to the people involved. In this case, the members of the community who underwent testing have a right to this information should they choose to have it.

Do Not Make the Combined Results Public

The knowledge that a certain proportion of the population carries a particular allele may cause individuals in the population to worry unnecessarily.

The data might make employers or insurance companies suspicious of all members of the community.

We should not disclose any potentially sensitive information without a persuasive reason to do so.

Figure T-4 Some reasons for and against making the combined results public.

and to ask questions about their significance. The following points may be helpful to you as you monitor and guide their discussion. If the students decide not to disclose the combined data, proceed immediately to Step 4 of the procedure.

□ Notice that the numbers listed on the overhead reflect only the number of *fictitious individuals* supposed by the imagined scenario to be carriers for each of these alleles. If you have assigned more than one group of students to any (or all) of the three families involved, the *actual number of your students* whose test results would show a positive result will be higher than the numbers on this form. For example, if two of your students (each from a different group) received the identity of Aisha Jacobs, then both of those students would see a positive result for cystic fibrosis if they opened their envelopes. The number "1" indicated on the transparency simply indicates that only 1 of the *fictitious people* who supposedly were tested (in this case, Aisha Jacobs) showed a positive result for this allele.

□ You also may want to point out to your students that the frequency of the allele for cystic fibrosis is 1/16 in the Caucasian population (the frequency is considerably lower in other populations) and the frequency of the allele for sickle cell disease is 1/20 in the African-American population (also considerably lower in other populations). The frequency of the allele for hypertrophic familial cardiomyopathy is difficult to determine because of its variable phenotypic expression and the relatively high rate of spontaneous mutation mentioned above.

Monitor your students' reactions to both the fictitious information and these actual frequencies. It is important that your students understand that potentially harmful alleles occur at frequencies that sometimes are surprisingly high and that many of us may bear one or more such alleles. This is one reason that each of us might reasonably be concerned about the privacy of genetic information. On the other hand, students should *not* finish the activities in this module feeling undue distress about their own genetic status. *If you are concerned about your students' responses, you might point out that to serve the teaching purposes of these materials, the three families depicted in the module show somewhat higher frequencies of deleterious alleles than one normally would expect in such a population.*

Disclosure Option 2
Report each set of individual results only to the person tested.

Introduce this option by saying something such as the following:

"Now we have to decide what to do with the individual results. Let's consider two options. First, I can give you the envelopes and each of you can open your envelope in private. In this case, we would release the individual results, but *only* to the people involved. How do you feel about this option? How should we proceed?"

List the reasons for and against such release as the students provide them. Figure T-5 provides some reasons *to release the results in this manner* and *not to release the results in this manner.*

Disclosure Option 3
Announce the individual results openly or publicly to all.

Introduce the discussion of the next option by saying something such as the following:

"A second option would be for me to announce the test results for each of you publicly, so each of us in the room will know everyone else's status with respect to the alleles in question. Is this acceptable? How do you feel about this level of disclosure? How should we proceed?"

Again, list the reasons for and against this level of disclosure as the students provide them. Figure T-6 provides some reasons *to announce the individual test results* and some reasons *not to announce the individual test results.*

Allow the students to decide by majority vote how to handle the individual data. Remind them that they will be choosing among three options for how to handle these data: disseminate them only to the individuals involved (Option 2); announce them publicly before the entire class (Option 3); or do not disseminate them at all.

Release to the Individuals Involved

Each person has a right and/or duty to know about his or her genetic status. Such information can help people make informed decisions about issues such as life style and reproduction.

This information is personal and sensitive, and each person should decide for himself or herself how—or if—the information should be shared.

No one else has any basic right to know this information.

This approach secures individual privacy. In fact, the potential for misuse indicates that information should go no farther than the individual.

Do Not Release to the Individuals Involved

Even this level of disclosure violates the privacy of other members of the family. For example, because Bob Thomas and Robb Thomas are identical twins, Bob's privacy is violated if Robb receives any information about his own genotype.

Completely private disclosure of this type is unwise. People would be receiving potentially upsetting information without the benefit of immediately accessible professional help, and they may seriously misunderstand or misapply that information. Should they choose not to seek professional help, they may carry private misconceptions about themselves that could alter decisions they make about their lives and reproductive behaviors. It may be better not to release the data at all than to do it in this manner.

This level of disclosure does not go far enough in ensuring good individual and societal outcomes. For example, individuals might choose to ignore important information about their health. As in the case of seat belt and helmet laws, it is the community's responsibility to assure that individuals do not act in ways that may be harmful to themselves. Because it is ultimately the community that pays for health care, it also might be argued that it is the community's right to have the information necessary to assure that each person acts in ways that will best protect his or her health and well-being. Without access to these data, the community cannot act in its own best interests. (This is an example of a case in which the rights of the community might override the rights of the individual.)

Other people who might be helped if they shared their test results (for example, their relatives and potential children) might not get the help they need if someone else does not persuade them to reveal this information. This is particularly relevant in the case of Bob and Robb Thomas. For example, if Bob Thomas tested positive for cardiomyopathy, but chose not to share this information with his brother, Robb would not have information that might save his life. This argument also could be extended to potential spouses, who may want to know about potential genetic risks within a family before making the decision to marry or to have children.

Completely private disclosure denies local health authorities, employers, insurance companies, and educational institutions the benefit of this information in making decisions.

The possible research benefits of this information also would be lost.

Figure T-5 Some reasons for and against releasing the information to the individuals involved.

Implement the class decision by distributing the envelopes for each person's private examination (Option 2), by opening the envelopes and reading the individual test results to the class (Option 3), or by discarding the envelopes containing the individual test results. If the class chooses Option 2 or 3, also display the overhead transparency that explains how students should interpret the test results (BLM T-6). Warn students that to avoid public disclosure of the test results, they should not remove the swabs from their envelopes. Instead, suggest that they carefully look inside the envelope to examine the swabs.

Disclosure Option 4
Enter the individual results into the LGD.

In the final portion of the activity, students consider a broader set of issues about the potential uses of genetic registries that contain genomic data. Because this module is focused on questions about the management, access, and regulation of genetic information in electronic databases, the activity is written to encourage more extensive discussion of this option than of earlier options.

3. Return to your small groups to consider a final option for disseminating the individual test results (Option 4) by answering the following question:

Should the individual test results from GeneTest, Inc. be entered into the LGD?

Use what you know about the genes in question, the arguments you and your classmates already have considered, and your experience with computerized databases as a basis for your discussion and decision.

Announce the Individual Test Results

All members of the class would know. This would alert class members, including teachers, to the potential risks of some activities for at least one individual in the class (the person carrying the allele for cardiomyopathy).

Once the data were made public in the class, the individuals involved likely would share it with other members of their families. This would allow all members of the extended family to make informed choices with respect to the potential for a genetic disorder in the family.

Likewise, as word spread, many individuals in the community likely would hear. This strategy would allow all members of the community to be aware of the importance of offering genetic services in the community.

Employers, insurance companies, local health authorities, and educational institutions will have better information about the genetic status of their potential employees, clients, and students. This will help them make better decisions about hiring, providing insurance, or planning health and educational programs. The community also could intervene if it becomes clear that individuals are behaving irresponsibly (for example, the school board might deny a student carrying the allele for cardiomyopathy the right to participate in athletic events).

Do Not Announce the Individual Test Results

This information is private and is no one else's business.

Once this information is made public, it will be very difficult to control how it is used. It might be used to discriminate against the individuals concerned in unfair ways.

This is not an appropriate approach to making sensitive information accessible to those who may have a reason to know it. This type of disclosure likely would lead to errors in both fact and interpretation among members of the community.

This would violate established practices concerning the handling of medical information by health-care practitioners, including genetic counselors.

Figure T-6 Some reasons for and against announcing the test results publicly.

Introduce Option 4 by saying something such as the following:

"Some of you might have voted against the public disclosure of your DNA test results because you were concerned about the informal, unrestricted, word-of-mouth communication suggested in Option 3. Remember that the LGD was organized in your town as a tool to help local health authorities collect important information about present and potential needs for genetic services in the community. What if we enter the test results into the LGD? Would this preserve the individual and societal benefits of disseminating these data, including the benefits to local health authorities, while avoiding the problems of the word-of-mouth spread of this important but possibly upsetting information?"

Allow 10 minutes for the students to list and clarify the reasons for answering either yes or no. We recommend that you assign each group the task of developing a list for *either* the "yes" or the "no" position. This will encourage some students to argue positions that they do not hold and may help all students clarify their thinking on the issues involved.

A good beginning point for discussion would be to identify the ways in which entering such data into a registry database is different from simply reading the names and test results to the class. Some of these differences are identified in the reasons *for* and *against entering the data into the LGD* that are provided in Figure T-7.

4. **Follow your teacher's instructions for sharing your list of "yes" or "no" reasons with the rest of the class.**

Following the brainstorming session, conduct a brief class discussion to elicit the "yes" and "no" reasons and post them on the chalkboard, the overhead, or on butcher block paper. Allow students to express their reasons completely. Prevent other students from interrupting to ensure that the most compelling reasons are articulated on both sides of the question. If necessary, add any reasons the students may have missed.

Enter the Data into the LGD

Entering the data into a registry provides a record of the actual test results that is more accurate than word-of-mouth communication. This disclosure option offers all of the benefits of the public announcement suggested in Option 3, without the dangers associated with informal, word-of-mouth transmission.

Entering the data into a registry database will assure its long-term availability and usefulness to research scientists, insurance companies, pharmaceutical companies, schools, and public-health officials both locally and around the world. This broad availability would benefit scientific research, would allow insurance companies and drug companies to develop and to market packages and products targeted to specific populations and individuals, would allow schools to respond more effectively to student needs, and would help public-health officials develop better ways to meet health-care needs in the area.

Do Not Enter the Data into the LGD

All of the arguments previously discussed in Options 2 and 3 for preserving privacy still apply. The potential for misuse of the data outweighs any benefit that might be gained by making the data publicly available.

The data that appear in a database may be in error, either as a result of an error in the sampling or testing process or as a result of a data-entry error. An error in the LGD may be more dangerous than those associated with word-of-mouth transmission of information because data in electronic databases often are viewed as more credible than other types of data. It also may be difficult for the people involved to get incorrect records changed.

Without stringent controls on access to the data in the LGD, the data could be downloaded by individuals or companies in any part of the world and might be used for purposes the community has not considered or authorized.

It is unwise to enter the data into the LGD in the absence of a clear and compelling reason to do so.

Figure T-7 Some reasons for and against entering the data into the LGD.

5. Consider both sets of reasons for a moment. What do you believe to be the most compelling reason or reasons on each side of the issue? Explain your answer.

Place a check mark beside the reason(s) the class decides is/are the most compelling on each side.

6. Are these two sets of reasons equally compelling? Why or why not? Explain your answer.

The students should see that one can cite ethically compelling reasons to support either decision (to enter the data or not to enter the data). This outcome, when it occurs, makes ethical decision making difficult. People who disagree in such circumstances should respect those with whom they disagree and not consider them to be persons of bad will. *The response to well-founded disagreement should be thoughtful respect and continued dialogue.*

7. Follow your teacher's instructions for deciding whether the individual test results should be entered into the LGD.

Allow the students to decide by majority vote whether to enter the data. If the students vote to enter the data, indicate to students that current medical practice probably would not permit entry of individual test results without the express consent of *each* individual. Even if the students vote not to enter the data, students should understand that the increasing use of computers in science and medicine suggests that each individual may not have full authority to decide where his or her private medical information is stored and how it is used. That is, students should understand that in the real world it is not likely that they would be given choices like this (although we may have some authority, we should not think that we have *full* authority to decide where such data reside and how they are used).

8. Can you think of any other situations in which access to personal genetic information in a registry might create ethical dilemmas?

Answers will vary. Suggest that the students think about the issues they encountered when using the LGD in Activities 1-3. For example, students might remember the questions that were raised in Activity 2 when a child's genotype was found to be incompatible with that of one or both parents and also about the dilemma that Joy faced as a result of the public dissemination of her genetic status in Activity 3.

Point out that the next activity will give the class an opportunity to consider some of the issues involved in making public policy with respect to the storage and release of genetic information from registry databases. You might note that as of 1995, there were no federal laws in place that expressly guarantee an individual's right to privacy with respect to genetic data, although various states have developed legislation to regulate the use of genetic information. In Activity 5, the students will act as state legislators who are faced with the responsibility of considering the relative merits of various policies about such databases. The materials that students read as homework will help prepare them for their deliberations and their vote.

Homework Assignment

In Activity 4, you considered a series of questions about how to handle the new genetic information that was gathered about your fictitious person as a result of additional DNA testing. Your discussions culminated in a decision about whether to enter those data into a registry database.

Your vote on this question likely was influenced by your experience with the databases in Activities 1-3. Those of you who voted to enter the data probably considered the advantages of having such information stored and accessible for research and health-care purposes. Those of you who voiced objections probably considered the disadvantages of having personal and potentially sensitive information accessible to those who should not have that information.

Many registry databases already exist in the United States. Some of these databases store actual genetic information; others store medical information from which one can infer

various types of genetic information. Many DNA databanks also exist. These are databases that store actual DNA samples or that store tissue samples from which DNA can be extracted. The individuals responsible for electronic databases and databanks usually adhere to relatively strict criteria about access to and use of these data. Many private citizens and organizations, however, believe that we need clearer and more uniform guidelines to protect the privacy of genetic information.

In Activity 5, your class will act as a committee of elected legislators to discuss and to recommend such guidelines. In preparation for your work as a legislator, read the background material on Worksheet 10, *Genetic Registries, Information, and Privacy*, and answer the study questions provided.

ANNOTATIONS TO WORKSHEET 10
Genetic Registries, Information, and Privacy

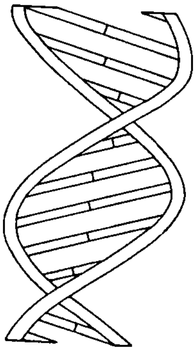
You may wish to point out to your students that Activity 5 begins in the same way that many state and federal policy deliberations begin—with the identification and the discussion of problems associated with one or more existing situations.

Reflect on what you learned from the activities that you already have completed and from the background information on Worksheet 10. Think as well about the PKU case.

a. What are some *advantages* of laws that place strict protections on the privacy of genetic information?

b. What are some *disadvantages* of such laws?

Answers will vary. Expect students to recognize that, in general, laws that place strict protections on the privacy of genetic information preserve individual autonomy (the right of each individual to be self-determining in thought and behavior) and limit the potential for misuse of this potentially sensitive information. On the other hand, such laws also may restrict society's ability to gain the beneficial consequences that can follow from the release of such information.



Activity 5
Making Public Policy

Focus

In this activity, students work as elected legislators to recommend a federal public policy related to the protection of genetic information that is stored in genetic registries and DNA databanks. Class discussion builds on the understandings that students gained during Activity 4 about ethical issues related to genetic information and also reflects the scientific and technical knowledge that students acquired about research databases and registries during Activities 1-3. By the end of Activity 5, students should understand that public policy is a mechanism for making decisions against a background of competing ethical positions. The students also should understand that *effective* policy anticipates and addresses likely as well as unlikely ethical concerns.

HGP Context

During Activity 4, the students discussed a series of questions about possible levels of disclosure of personal genetic information and about the storage and use of information in a genetic registry. This experience gave students the opportunity to consider how to balance competing ethical commitments and to discover that it is not always easy to decide, once and for all, among various courses of action.

Public policy serves an important function here: to clarify and to advance certain decisions and actions against a background of debate. The question of what policies to establish—a question that will face all of us as the HGP progresses—is really a question of how public policy should respond when policy makers cannot advance all of the interests of all of the affected parties. Particular policies can advance the interests of some people, but only at the price of impairing the interests of others. As students work through the arguments in Activity 5, they should begin to understand that public policy is a mechanism through which self-governing people manage situations in which not all interests can be advanced.

Major Concepts

□ Ethics brings to public-policy debates about genetic registries two important presumptions: first, that we should protect individual autonomy, and second, that we should protect individual and societal health and well-being.

□ The presumption that we should protect individual autonomy serves as a powerful ethical justification for the right to privacy and the duty of confidentiality. This presumption provides moral grounding for policies aimed at the practice of securing informed consent in the health-care setting.

□ The presumption that we should protect individual and societal health and well-being serves as a powerful ethical justification for paternalistic intervention to protect individual health, especially of children. This presumption has deep roots in ethical theories that are concerned with promoting human good as a highly valued consequence.

□ Sometimes it is possible to propose ethically justified trade-offs between policies that favor autonomy and policies that favor individual and societal health and well-being. The challenge in such a case is to develop criteria that can clearly and prudently inform us about when autonomy or privacy can be overridden.

Student Objectives

As students complete this activity, they should

□ understand that the analysis of ethical issues related to accessing and using the information stored in genetic registries highlights two important values, the value of autonomy and the value of individual and societal health and well-being;

□ understand that these values sometimes conflict in public-policy debates about genetic registries; and

□ explain that society can respond to ethical issues about genetic registries in one of three ways: (1) by enacting laws that favor autonomy as the overriding consideration (for example, laws that protect the privacy and confidentiality of genetic information stored in registry databases); (2) by enacting laws that favor individual and societal health and well-being as the overriding consideration (for example, laws that allow privacy and confidentiality to be violated when doing so would gain highly valued consequences for individuals or for society); or (3) by not enacting laws, a strategy that allows more time for public discussion and debate.

Science Process Skills

□ Synthesizing information and knowledge

Public-Policy Process Skills

□ Analyzing issues

□ Evaluating issues

□ Communicating (orally and in writing)

Materials for a Class of 30

□ 30 sets of student text, pp. S-35–S-36. Because students will complete their written work on their own paper or on the worksheets provided, students can reuse these pages.

□ 10 copies each of Worksheet 11, *Analyzing Recommendation A*; Worksheet 12, *Analyzing Recommendation B*; and Worksheet 13, *Analyzing Recommendation C*.

□ 1 copy of Worksheet 14, *Instructions for the Committee Chairperson*.

□ Overhead transparency of the analysis chart for the three recommendations under consideration (BLM T-7).

□ Name tags for the chairperson and members of the legislative committee (optional).

Advance Preparation

□ Make the necessary student copies and overhead transparency.

□ If desired, make name tags that will identify a student as the chair of the legislative committee and that will identify each member of your class as a member of that committee. You may wish to indicate on these tags the particular subcommittee (A, B, or C) on which each student will serve.

Estimated Time

One 45-minute class period

Introduction

The underlying assumption of Activities 4 and 5 is that ethics and public policy are integrated ways of thinking about what human activities and behavior should or should not be promoted. This relationship between ethics and public policy should become apparent to your students as they move from Activity 4 to Activity 5. Specifically, the students should see that their discussions in Activity 4 equipped them with a number of useful conceptual tools for the public-policy issues that they consider in Activity 5.

For example, although the students may not express their understanding in formal terms, they should see that ethics brings to public-policy debates about registry databases two important presumptions: first, the presumption that we should protect individual autonomy, and second, the presumption that we should protect individual and societal health and well-being. These values (or interests) clearly are at stake for individuals, institutions, and society in the PKU case that students read about as part of their homework following Activity 4. On the one hand, the insurance company can claim that it has a right to pursue its own interests, in this case, to engage in business in the free market as it chooses. In addition, administrators of the insurance company might point out that taking on too many high-risk individuals would affect the company adversely. These officials might argue that to protect the company from such a situation (specifically, to protect the company against what insurance companies call *adverse selection,* individuals seeking insurance without providing full and accurate information about their medical backgrounds), the company must have the right to acquire and use relevant medical information. The company also might argue that it has a right to pool its information resources with other insurance companies (in a central information system) and to access and use this information as it deems appropriate.

In contrast, the child's family might argue that there is no real conflict between the family's interests (gaining health-care coverage for their daughter) and the company's interests. In support of this position, the family might point out that the insurance company is basing its contention of unreasonable risk on an improper understanding of the information that it obtained. For example, the child's normal development to date might suggest that her risk for illness is no greater than that for other children her age. Of course, it is appropriate to recognize that the child would require regular visits to the clinic to have her phenylalanine levels monitored and to receive dietary counseling. The family also might complain that the insurance company's decision reflects an inadequate understanding of its responsibility to share risks and costs among its clients. Put another way, *cherry-picking,* that is, choosing the most desirable and lowest-risk clients, violates the normal mission of an insurance company. Clearly, it is in the daughter's interests to gain health-care coverage (that is, to share her risk of significant health-care costs with others). From the family's perspective, the tension between her interests and those of the insurance company (which sells risk-sharing) may hinge entirely on how much risk each party thinks the company should bear. One also might claim that the daughter and the family have a right to health care, or at least a right not to be harmed by the inappropriate use of private information. In any case, the question of fair access and appropriate use of the genetic information stored in registry databases is central to the problem.

In this activity, we use the PKU case to begin deliberations about making good public policy. It is important to recognize, however, that good public policy comes about through deliberations involving *numerous* cases and issues. Nevertheless, policy making often begins with a prominent case or two. It is at this initial point of policy making that the students begin their work.

Activity 5 offers students the opportunity to discover that a key difficulty in developing effective public policy is that any single policy typically advances one set of interests over the other: in the PKU case, *either* the insurance company's interests *or* the family's interests. At the end of the activity, students should see the challenges of developing good public policy. In addition, students should understand that effective policies balance competing ethical positions in ways that most people see as justified.

Related Information in the Teacher Narrative

You will find additional background information in the following pages of the teacher narrative:

□ registry databases in the United States (pp. 25-28)
□ legal protections of privacy of genetic information in the United States (pp. 43–44)
□ ethical issues related to registry databases (pp. 36–40)
□ teaching about values and controversial issues (pp. 44–45)

Annotated Student Activity

Imagine that you are an elected federal legislator and a member of the legislature's Committee on Genetic Information in Registry Databases. Your first task is to elect a student to serve as the committee chairperson; this individual will guide today's discussion and conduct the concluding committee vote.

Help the class elect a classmate to serve as the committee chairperson. This individual should call the committee to order at the end of the discussion period, guide the subcommittee reports, and call the final vote.

Three of your legislative colleagues who are not on the committee have submitted separate recommendations for your consideration. These recommendations are listed below.

Recommendation A:

It is premature to act. There should be no *new* laws developed at this time about the release or the use of genetic information from registry databases.

Recommendation B:

There should be laws that limit an individual's right to privacy with respect to the data in genetic registries when release of the data can be shown to be in the best interests either of that individual or of the community as a whole.

Recommendation C:

There should be laws that guarantee an individual's right to privacy with respect to the data in genetic registries.

You may wish to point out to your students that Recommendations A, B, and C form a continuum

□ from the present situation, in which few formal protections to privacy exist and, among those that exist, there are inconsistencies across states (Recommendation A);

□ to a situation in which formal protections would apply to genetic information stored in all genetic registries, although these protections could be overridden under certain circumstances (Recommendation B); and

□ to the last situation, in which these protections to privacy would be absolute, with no opportunity for relaxation, even under unusual cases (Recommendation C).

To ensure that each of these recommendations is considered carefully, the committee will form three subcommittees. Each subcommittee will discuss one of the recommendations, focusing particularly on how policies consistent with that recommendation would work in the case that you considered in preparation for these discussions. After analyzing its assigned recommendation, each subcommittee will report its findings to the full committee. The committee then will decide by majority vote which recommendation it will present to the legislature for further action.

Procedure

1. Follow your teacher's instructions for meeting in your subcommittee.

To encourage maximum participation by all students, you may wish to establish six small groups rather than three large groups. If you do so, simply assign each recommendation to two different groups.

You may choose to allow students to decide which subcommittee they serve on (that is, which recommendation they analyze), or you may choose to ask students to remain in the same groups they worked in for the preceding activities and assign each group a specific recommendation to anzlyze. An advantage of the first option is that students may have specific reasons for wanting to consider particular recommendations; an advantage of the second option is that it may lead to more balanced representation on each subcommittee and it may require some students to consider carefully a recommendation that they do not initially understand or support.

2. Your teacher will provide you with a worksheet that states the recommendation that

your subcommittee will discuss and that outlines the questions that you will be asked to answer when you report to the full committee. Use the background information on current privacy laws in the United States, the case, and your answers to the study questions that you completed for homework as resources to help you analyze your assigned recommendation. You will have about 15 minutes to complete your deliberations.

Distribute one copy of the appropriate worksheet (Worksheet 11, 12, or 13) to each student serving on a subcommittee. Give the students approximately 15 minutes to discuss the recommendations and complete their analyses.

Give one copy of Worksheet 14 to the student chairperson. This person should study the information provided on the worksheet so that he or she will be prepared to lead the upcoming discussion.

3. Choose a spokesperson to report the results of your subcommittee's analysis to the full committee. Follow the committee chairperson's instructions for completing and discussing the reports from the subcommittees.

Use BLM T-7 to help the chairperson organize and summarize each subcommittee's report.

Encourage the chairperson to stimulate discussion of each subcommittee report by pausing after the report is complete and asking the class whether anyone has questions that they would like to ask about that recommendation or if anyone would like to raise other issues or arguments that the subcommittee did not appear to consider.

4. Remember that you, as well as your colleagues, eventually will be asked to vote for the recommendation that you believe is most likely to lead to actions that will best address questions about the protection of data in genetic registries. In preparation for voting, follow your chairperson's instructions for participating in a general discussion of all of the recommendations. The questions below will help you organize your thinking about this issue.

□ Which recommendation, if any, do you think works well in the PKU case? Explain your reasoning.

□ Which recommendation, if any, do you think would work well in most cases? Explain your reasoning.

□ What questions do you have about these recommendations that have not been raised or addressed? What alternative recommendations do you have to offer?

The chairperson should invite a general discussion of the relative merits of each of the recommendations at this time. Make sure that the students understand that all three of the recommendations represent a decision to do *something* (even Recommendation A is a decision to let the current situation continue).

Students may find it difficult to articulate clear alternatives to the recommendations provided. Help the students see that there is a range of responses that they might have to these recommendations. Some students may feel that they are sufficiently comfortable with one of the recommendations that they see no need to suggest an alternative or to modify, refine, or further elaborate any of the others. Other students may feel that one or another of the recommendations would be more acceptable to them if it were modified or more clearly delineated. For example, some students may suggest modifying Recommendation B to include a list of guidelines for deciding when an individual's privacy could be overridden. Finally, some students may see completely new alternatives that the committee has not yet considered.

Encourage students to share their own responses and also to reflect and comment on questions and alternatives that other students raise. Many students may find this discussion very challenging. Assure them that this is not unusual: policy makers at both federal and state levels also find these issues complex and difficult to analyze.

It may be the case that students will remind each other that policy usually is recommended on the basis of a number of cases. If this discussion occurs, ask students to think about whether their support for any one of the policies would be changed or altered if Jennifer were diagnosed with a more serious condition that would require a great deal of clinical intervention and, in turn, would increase costs to the insurance company (and, potentially, to other policyholders).

5. **Follow your chairperson's instructions for casting your vote.**

NOTE: You are *not* required to vote for the recommendation that you analyzed. Instead, you should vote for the recommendation that you think would lead to the best outcome.

The chairperson should ask the students to vote on which recommendation to propose to the legislature for further action.

6. **Reasonable people can come to very different decisions about a controversial question or issue. What does this suggest about the process of developing public policy?**

You may wish to organize this discussion. This outcome highlights the complexity, but also the importance of developing public policy. Students should recognize that particular policies advance certain interests and impair others. The process of developing public policies helps society clarify its thinking on controversial issues and allows self-governing people to manage situations in which not all interests can be advanced.

Annotations to Worksheet 11
Analyzing Recommendation A

Recommendation A states that it is premature to act at this point and recommends that there be no *new* laws developed at this time about the release or the use of genetic information from registry databases. The effect of such a recommendation is to allow more time for public debate and discussion. When matters are unsettled, it is important that this debate take into account the interests of all affected individuals, communities, and institutions. A way to gain the time required for this type of discussion is to rely temporarily on *de facto* public policy. Essentially, this means that whether or not access to genetic information would be allowed in each of these cases depends on the specifics of the case, as well as on the context in which it occurs.

It is important that your students understand that this does not mean that in the absence of a new law, anything goes. Rather, it means that issues would continue to be resolved in the same way that they have been to this point. That is, if there already is a law that covers a particular situation, then, in the absence of a new law, the current statutes stand. Issues that have not been legislated already would remain unregulated, in the absence of a new law. Even this, however, does not mean that those who control the data in question are free to act irresponsibly in decisions about its access. Men and women do share certain common ways of behaving and, even in the absence of specific law, those behaviors tend to continue.

Consider the PKU case. Would this recommendation have allowed the insurance company to access and use the data in this manner? (circle one) yes no

Yes; this recommendation would maintain the current status of genetic information. Under these conditions, the insurance company would have been free to access and use the information in this manner.

Give some reasons that this would be a bad outcome. Which reason do you think is most important? (check one)

The current legal situation allowed this insurance company to access information that led to the denial of health-care insurance to this child and her family. If we allow this situation to continue, other individuals and families will find themselves facing similar problems. Fear of how genetic information might be stored and then used against them or against members of their families might lead individuals to feel trapped in their current jobs (to change jobs risks re-evaluation by a new health-care provider), and might even lead people to refuse to seek genetic testing or treatment for various genetic conditions, even when such testing or treatment could bring immediate benefit to them.

Give some reasons that this would be a good outcome. Which reason do you think is most important? (check one)

Insurance companies need this type of information to measure accurately the risk of providing coverage to specific individuals or families. If insurance companies were denied this information (effectively forcing them to provide equal coverage to everyone, without knowing their general state of health or their genetic status), the costs of insurance for all of us might increase. This argument is even stronger if the insurance company providing the group coverage for the father's new employer is basing its rates on past experience of the company's collective health-care costs. If an individual or a family with higher than average health-care costs is added to this group, the effect may be to drive the costs for the group plan higher, adversely affecting all of the employees in the company and company profits.

Discuss the two reasons that you have identified as being the most important. Place a second check in front of the reason that your subcommittee finds the most compelling.

Do you think that Recommendation A works well in *the PKU* case? (circle one) yes no

Answers will vary, depending on the relative importance that the group has assigned to the arguments they have identified.

Do you think that Recommendation A would work well in most cases? Explain your answer and support it with other specific examples of situations in which you think that it would or would not work well.

Again, expect students to offer different answers depending on how concerned the group is about the issues involved.

Annotations to Worksheet 12
Analyzing Recommendation B

Recommendation B proposes that there should be laws that limit an individual's right to privacy when the release of the data in question can be shown to be in the best interests either of that individual or of the community as a whole. The goal of Recommendation B, then, is to protect individual and societal health and well-being.

Consider the PKU case. Would this recommendation have allowed the insurance company to access and use the data in this manner? (circle one) yes no

Whether laws consistent with this recommendation would allow access to and use of the data in this case would depend on the criteria that these laws established for determining the well-being or "best interests" of the individual or the community. In the absence of specific information as to what these criteria might be, it is impossible to know whether the insurance company would continue to have access to such data.

At first glance, Recommendation B may seem to students to be an appropriate and moderate way to resolve the tensions between the values of privacy and autonomy on the one hand and the values of individual and societal health and well-being on the other. *As the students begin to examine the recommendation more closely, however, they should see that the success of such laws depends on the criteria used to determine when privacy can be overridden. Put another way, much depends on the criteria used to determine when the best interests of the relevant parties are being promoted or defeated.*

Give some reasons that this would be a bad outcome. Which reason do you think is the most important? (check one)

The students may answer this question differently, depending on how they resolved the preceding question. If they recognize that they cannot provide definite answers to questions about outcomes because they do not know whether the data would be available to the insurance company, suggest that they examine both possibilities: the possibility that the criteria that were eventually established *would* allow use and the possibility that the criteria *would not* allow such use.

Refer to the suggested answers on Worksheets 11 and 13 for ideas about how students might evaluate the outcomes in this situation.

Give some reasons that this would be a good outcome. Which reason do you think is the most important? (check one)

Again, the students may answer this question differently depending on how they resolved the first question on the worksheet. Refer to the suggested answers on Worksheets 11 and 13 for ideas about how students might evaluate the outcomes in this situation.

Discuss the two reasons that you have identified as being the most important. Place a second check in front of the reason that your subcommittee finds the most compelling.

Do you think that Recommendation B works well in the PKU case? (circle one) yes no

Answers will vary, depending on the relative importance that the group has assigned to the arguments they have identified.

Do you think that Recommendation B would work well in most cases? Explain your answer and support it with other specific examples of situations in which you think that it would or would not work well.

Expect students to offer different answers depending on the relative importance that the group has assigned to arguments related to privacy and arguments related to issues of individual and societal health and well-being.

ANNOTATIONS TO WORKSHEET 13
Analyzing Recommendation C

Recommendation C states that there should be laws that guarantee absolutely an individual's right to privacy with respect to the data in genetic registries. The goal of Recommendation C is to protect privacy.

Consider the PKU case. Would this recommendation have allowed the insurance company to access and use the data in this manner? (circle one) yes no

If laws guaranteeing an individual's absolute right to privacy were passed, it is possible that this level of access to genetic information would be denied to insurance companies.

Give some reasons that this would be a bad outcome. Which reason do you think is the most important? (check one)

This would be a bad outcome because insurance companies need this type of information to measure accurately the risk of providing coverage to specific individuals or families. If insurance companies were denied this information (effectively forcing them to provide equal coverage to everyone, without knowing their general state of health or their genetic status), the costs of insurance for all of us may increase. This argument is even stronger if the insurance company providing the group coverage for the father's new employer is basing its rates on past experience of the company's collective health-care costs. If an individual or a family with higher than average health-care costs is added to this group, the effect may be to drive the costs for the group plan higher, which could adversely affect all of the employees in the company.

Give some reasons that this would be a good outcome. Which reason do you think is the most important? (check one)

This would avoid the situation in which this insurance company would be able to access information that led to the denial of health-care insurance to this child. If we deny such companies access to this type of information, we protect other individuals and families from facing similar problems. This outcome also might reduce the fear that genetic information will be stored and then used against individuals or against members of their families. This would help people feel more autonomous and might encourage more people to seek genetic testing or treatment for various genetic conditions, especially when such testing or treatment could bring immediate benefit to them.

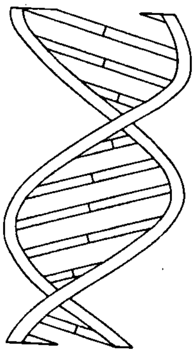
Discuss the two reasons that you have identified as being the most important. Place a second check in front of the reason that your subcommittee finds the most compelling.

Do you think that Recommendation C works well in the PKU case? (circle one) yes no

Answers will vary depending on the relative importance that the group has assigned to the arguments they have identified.

Do you think that Recommendation C would work well in most cases? Explain your answer and support it with other specific examples of situations in which you think that it would or would not work well.

Again, expect students to offer different answers depending on the relative importance that the group has assigned to arguments related to privacy and arguments related to issues of individual and societal health and well-being.



Extension Activity HGP Data and Evolutionary Biology

Focus

This activity introduces students to the use of genomic databases in the study of evolution. Part A highlights the use of DNA data to establish relationships between members of the same species, in this case *Homo sapiens*. Part B focuses on the use of DNA and amino acid sequence data to establish evolutionary relationships between different species.

HGP Context

Evolution is the central organizing concept of biology and the HGP, the largest single project in the history of the biological sciences, likely will shed considerable light on the mechanisms of evolution. During the last 20 years, scientists from a variety of disciplines increasingly have brought data from molecular biology to bear on the study of evolutionary relationships within and among species and higher taxonomic groups.

The HGP supports research on organisms other than *Homo sapiens*, and sequence data from these organisms allow us to analyze evolutionary relationships. Such research, however, would be extremely difficult without large sequence databases, and an important part of the HGP is improvement of information technologies that promote the effective storage and manipulation of large amounts of sequence data. These data ultimately will improve our understanding of general phylogeny as well as our knowledge of human history, racial relationships, and migration patterns.

Major Concepts (Part A)

□ We can establish relationships between individuals by using DNA data.

□ Genomic databases allow us to ask questions about evolution that we could not ask before those data were available in a form that was readily accessible.

Major Concepts (Part B)

□ All organisms on earth are related through evolution.

□ The evolutionary history of a species is written in its DNA.

□ Life begets life; therefore, relationships between individual species are a function of descent with modification.

□ We can establish relationships between different species by using DNA sequence data and amino acid sequence data.

□ Amino acid sequence data allow us to look further back in time than do DNA sequence data.

□ The volume of the data in studies of molecular evolution requires the use of computers.

□ Genomic databases allow us to ask questions about evolution that we could not ask before those data were available in a form that was readily accessible.

Student Objectives

As students complete this activity, they should

□ understand that differences in DNA can be used to establish relationships between individuals of the same species and between members of different species;

□ manipulate sequence data to establish the degree of identity between sequences;

□ use sequence data to infer relationships between members of the same species (*Homo sapiens*);

□ construct a simple phylogeny on the basis of amino acid sequences;

□ understand that sequence databases provide resources for studying evolutionary relationships; and

□ understand that the HGP will provide vast amounts of data from humans and other organisms for use in research on evolution.

Science Process Skills

□ Analyzing and manipulating data

□ Using the computer as a research tool

□ Proposing hypotheses to explain observations

□ Revising hypotheses based on new data

□ Drawing conclusions based on analysis of data

Materials for a Class of 30

□ 30 sets of student text, pp. S-43–S-46. Because students will complete their written work on their own paper or on the worksheets provided, students can reuse these pages.

□ 30 copies each of Worksheet 15, *Matching DNA Sequences from Frog, Chicken, Goat, Cow, and Chimpanzee*; Worksheet 16, *DNA Sequence Data for the Beta Hemoglobin Gene in Humans*; Worksheet 17, *Constructing an Evolutionary Tree from DNA Sequence Data*; and Worksheet 18, *Constructing an Evolutionary Tree from Amino Acid Sequence Data.*

□ Overhead transparencies of BLM T-8, *Numbers of Differences in DNA Sequences*, and BLM T-9, *A Simple Evolutionary Tree.*

Advance Preparation

□ If you have not done so already, load the software onto computers. Set the software on each computer to *HGP Data and Evolutionary Biology, Part A*. You will find instructions for how to do this on p. 56. The password is <Mendel>.

□ Make the necessary student copies and overhead transparencies.

Estimated Time (Part A)

One 45-minute class period

Estimated Time (Part B)

One 45-minute class period

Annotated Student Activity

Part A
DNA, Political Assassination, and World History

Genomic databases can include DNA sequence data from living individuals or from persons long dead. The sequence data for this activity, which you will retrieve from the research database, include actual data from DNA taken from nine human skeletons found in a shallow grave in 1991.

The DNA used for this analysis is mitochondrial DNA, that is, DNA taken from the mitochondria of cells. Recall from your study of cells and of energy production in living systems that mitochondria are contained in the cytoplasm of the cell. Their primary function is the production of ATP, the chief source of energy in living systems. Mitochondria have their own DNA, apart from the DNA found in the nucleus. The complete sequence of human mitochondrial DNA, or mtDNA, as it is called, was determined in 1991. The sequence of 16,569 bases is included as part of international DNA sequence databases.

If your students are not familiar with current hypotheses that explain the existence of a separate mitochondrial genome, summarize as follows:

Mitochondria may have originated as bacteria—complete with a bacterial genome—that parasitized cells. This parasitism ultimately benefited host and parasite, providing selective advantages to both. In other words, the parasitism became symbiosis/mutualism, and the bacteria/mitochondria became an integral part of cell structure and function.

All human cells—including ova and sperm—contain mitochondria. Biologists have observed, however, that mitochondria are transmitted only through females—to their sons and daughters. Propose a hypothesis to explain this observation.

HINT: Consider the location of mitochondria in sperm cells. What happens to a sperm cell during fertilization?

Mitochondria are concentrated in the sperm's tail region, providing energy for locomotion. Only the head of the sperm penetrates the ovum during fertilization. All mitochondria in the zygote, therefore, are of maternal origin.

In addition to being maternally inherited, some regions of the mitochondrial genome mutate readily, so there is enough variation to distinguish families from one another.

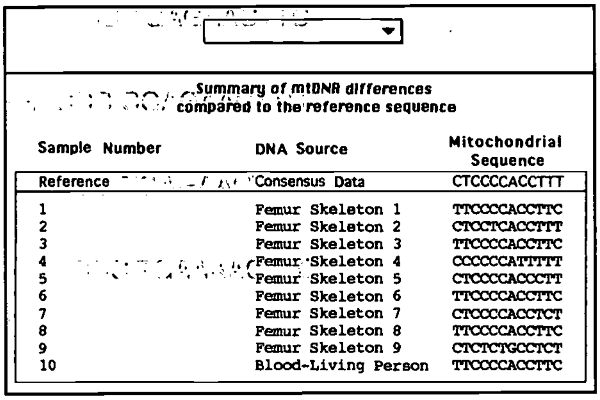
PROCEDURE

Part A is designed to be completed in one 45-minute class period. To activate the data required for this activity, select *HGP Data and Evolutionary Biology, Part A* from the *ACTIVITY* pull-down menu. You may either perform this function on each of your students' computers or ask them to complete it. The password is <Mendel>.

1. The first screen for this activity is shown in Figure 7. Note that it displays the following:

- **□ reference data for a 12-base sequence of human mtDNA,**
- **□ mtDNA sequences from nine skeletons recovered from a shallow grave in 1991, and**
- **□ an mtDNA sequence from a living person.**

We treat these data as a contiguous sequence, although in reality they are the bases from 12



Summary of mtDNA differences compared to the reference sequence

Sample Number	DNA Source	Mitochondrial Sequence
Reference	Consensus Data	CTCCCCACCTTT
1	Femur Skeleton 1	TTCCCCACCTTC
2	Femur Skeleton 2	CTCCTCACCTTT
3	Femur Skeleton 3	TTCCCCACCTTC
4	Femur Skeleton 4	CCCCCCATTTTT
5	Femur Skeleton 5	CTCCCCACCCTT
6	Femur Skeleton 6	TTCCCCACCTTC
7	Femur Skeleton 7	CTCCCCACCTCT
8	Femur Skeleton 8	TTCCCCACCTTC
9	Femur Skeleton 9	CTCTCTGCCTCT
10	Blood-Living Person	TTCCCCACCTTC

Figure 7 Sample screen from *HGP Data and Evolutionary Biology, Part A.*

different regions of the mitochondrial genome. These regions are useful because of their variability.

2. **To make it easier to find sequences that are identical, it is convenient to use a dot (.) to replace any base that is identical to the base that occurs at the same position in the reference sequence. Bases that do not match remain listed as letter symbols. Follow the instructions below to insert dots in any positions in which the bases are identical to those in the reference sequence. What do you notice about the sequences?**

Five of the sequences are identical. The use of dots is only one of many styles used in the biological literature to display sequence data more conveniently.

Note that students can toggle back and forth between the screen that shows all bases and the screen that replaces identical bases with dots simply by highlighting the appropriate phrase (*Replace with Dots* or *Begin Again*) in the pull-down menu. Encourage them to use this feature to see that all bases identical to those in the reference sequence have, in fact, been replaced by dots.

To Replace Identical Bases with Dots

Click and pull on the downward-pointing triangle that appears in the box at the top of the screen. Highlight the line that reads *Replace with Dots* and release the mouse button. The computer will replace any base that is identical to the reference data with a dot.

3. **Work with your team members to cluster the sequences that are most similar.**

NOTE: You cannot move the reference sequence.

Impress upon the students that they should manipulate the sequences until they have maximized the similarities.

4. **Among the recovered skeletons are those of an adult female and three children. Based on this information, the sequence data, and the method of transmission of mtDNA, propose some hypotheses about the skeletal remains.**

To Move the Sequences

For Macintosh Computers

- Click and drag the sequence you wish to move. Release the mouse when the cursor reaches the desired position.

NOTE: The sequence will not appear to move at first. Instead, an arrow will appear to the left of the sequence you are dragging. This arrow indicates that the sequence *will* move when you release the mouse button.

For Windows Computers

- Click on the sequence you wish to move. This will highlight it.
- Click on the UP or DOWN button to move the highlighted sequence to the desired position.

Hypotheses include the following:

□ the three children are the children of the adult female, and

□ the fifth identical sample (blood sample from a living person) is from a living relative of the mother and her three children.

5. **As a class, read and discuss the following information:**

The official records of the fate of the Russian royal family (the Romanovs) are scant, but it is known that prior to their deaths they were held prisoner in Ipatiev House at Ekaterinburg, in the Ural Mountains of Central Russia (Figure 8). It is believed that shortly after the night of 16 July 1918, Tsar Nicholas II; his wife, Tsarina Alexandra; their four daughters, Olga, Tatyana, Maria, and Anastasia; and their only son, Alexei, were herded into the cellar together with three of their servants and the family doctor, Eugeny Botkin. They were shot by the Bolshevik firing squad, although a number of the victims were allegedly stabbed to death when gunfire failed to kill them. The bodies were stripped and placed onto a truck with the intention of disposing of them down a mine shaft. However, the truck developed a mechanical fault during the journey; the Bolsheviks placed the victims into a hastily dug pit in a road, and to hinder identification,

sulfuric acid was thrown into the open grave. After the bodies were covered, a truck was driven backwards and forwards over the site to flatten the area.

This account is supported by forensic evidence collected in 1918–19 by Nikolai Sokolov (a White Russian monarchist investigator) whose seven-volume dossier has become the basis for the historical version of the fate of the Romanovs. However, the version of events described above has never been positively verified.

After referring to archival materials and photographs, which gave an indication of a burial site, two Russian amateur historian investigators, Gely Ryabov and Alexander Avdonin, announced that they had discovered a communal grave approximately 20 miles from Ekaterinburg. Consequently, the Russian government authorized an official investigation coordinated by the chief forensic medical examiner of the Russian Federation. The grave consisted of a shallow pit (less than 1 m deep) and contained human skeletal remains. Many of the bones were badly damaged, but it was possible to identify nine corpses. All of the skeletons showed evidence of violence and mistreatment before death. Some of the skulls had bullet wounds; bayonet marks were also found. Facial areas of the skulls were destroyed, rendering classical facial identification techniques difficult.[4]

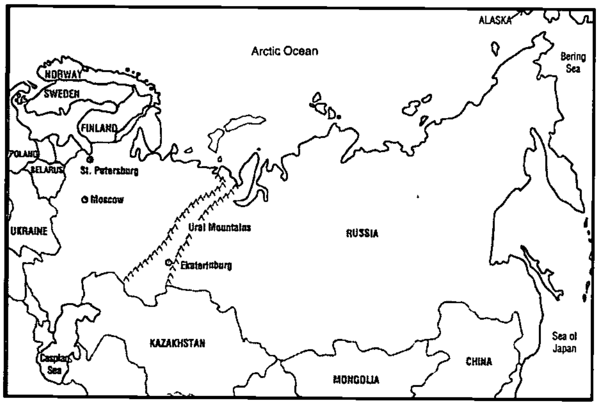


Figure 8 Map of the former Soviet Union.

For students interested in classic rock, point out that this famous and critical event in world history is part of the song *Sympathy for the Devil,* by the Rolling Stones:

> "Stuck around St. Petersburg,
> When I saw it was time for a change.
> Killed the Tsar and his ministers,
> Anastasia screamed in rage."

This is an excellent opportunity to involve your social studies faculty in discussions of Russian and world history, emphasizing the Bolshevik revolution and the development and demise of European communism.

Now tell your students the following: "The mtDNA data you analyzed are the actual data from nine skeletons found in that shallow grave about 20 miles from Ekaterinburg, Russia, in addition to mtDNA from the blood of a living person. Would you like to propose any additional hypotheses about the skeletons or about the living person?"

Students likely will propose that the female skeleton is that of Tsarina Alexandra, and that the three identical remains are those of three of her five children. They also may propose that the fifth identical sample is from a living relative of Alexandra.

6. Follow the instructions below to evaluate the clusters and discover the origin of the samples. Do these data support your hypotheses? Do the data prove your hypotheses?

Note that the sequences may be listed in a different order on each group's screen, depending on the manner in which the students chose to cluster the samples (for example, with identical sequences at the top or the bottom). Regardless of the particular order, the computer will identify each sample properly.

To Evaluate the Clusters

Click and pull on the downward-pointing triangle that appears in the box at the top of the screen. Highlight the line that reads *Evaluate Clusters* and release the mouse button.

The data support the hypothesis concerning the Tsarina and her children. The blood sample comes from His Royal Highness (HRH) Prince Philip, the current Duke of Edinburg (see the pedigree in Figure T-8). The Duke is a grandnephew of the Tsarina, through his grandmother, Victoria of Hesse, the Tsarina's sister. The DNA data do not prove these relationships conclusively. The scientists who conducted this work, however, conclude that the current odds are at least 700 to 1 that they are correct in their analysis. They expect those odds to increase as data on mutation rates in mtDNA improve. This is a good opportunity to discuss the difference between support and proof in science.

The skeletons found in Ekaterinburg included the remains of only three children—all female. The grave contained no signs of the Tsar's and Tsarina's only son, Prince Alexei, nor of their fourth daughter. This fourth daughter, the youngest, is believed to be Grand Duchess Anastasia, the subject of a decades-long search and numerous false claims of identity.

One woman who claimed to be Anastasia was Anna Anderson, of Charlottesville, Virginia, who died in 1984. The Russian Nobility Association disputed Anderson's claim, but others who knew her were convinced that she was, indeed, Anastasia.

Anderson first drew attention with her claims of nobility as a young patient in a Berlin mental hospital in 1921. In the late 1950s, before leaving Germany for Virginia, Anderson had provided

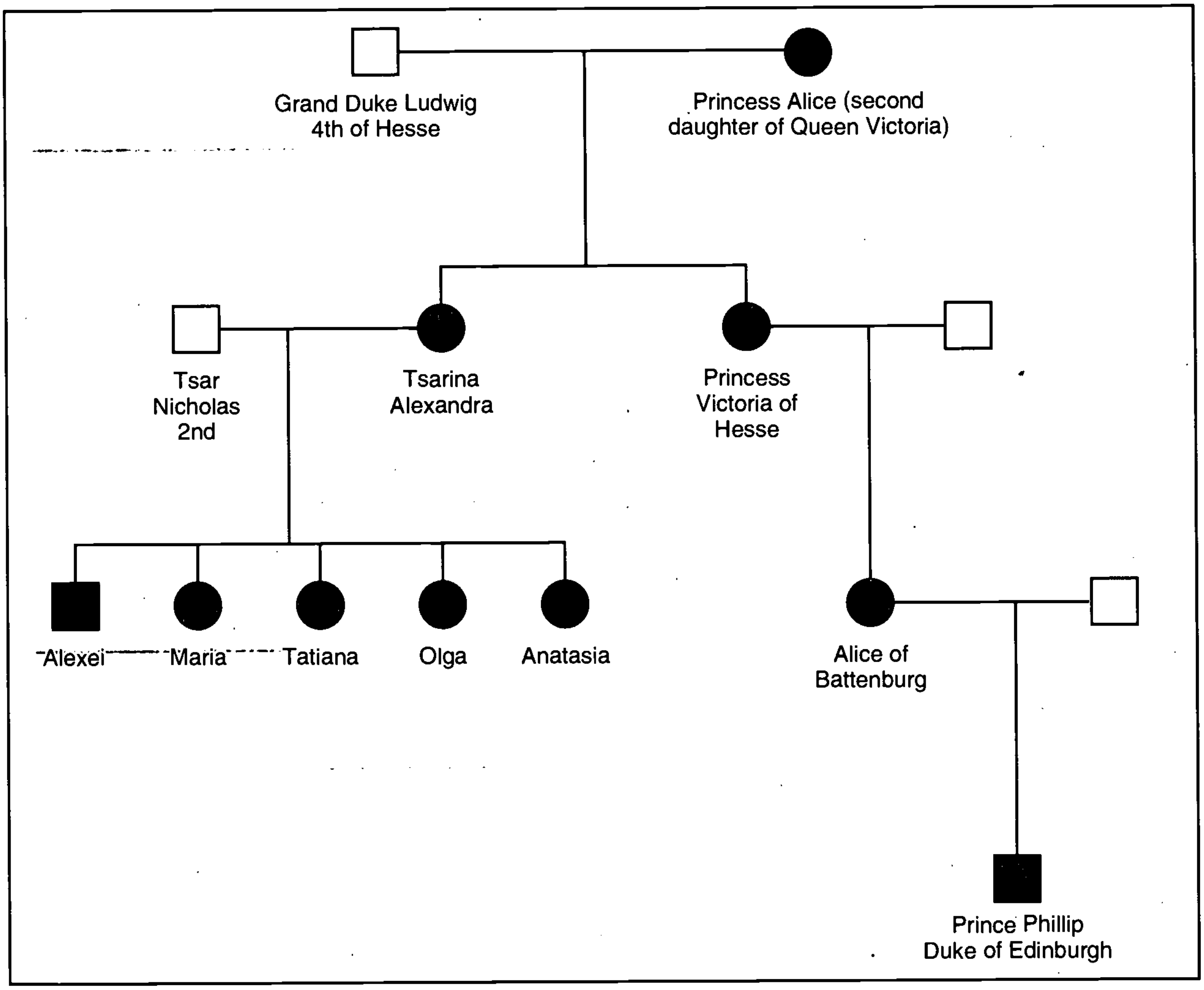

Grand Duke Ludwig 4th of Hesse
Princess Alice (second daughter of Queen Victoria)
Tsar Nicholas 2nd
Tsarina Alexandra
Princess Victoria of Hesse
Alexei
Maria
Tatiana
Olga
Anatasia
Alice of Battenburg
Prince Phillip Duke of Edinburgh

Figure T-8 Lineage of Tsarina Alexandra showing relationship to HRH Prince Philip (Duke of Edinburgh). Darkened symbols indicate a maternal lineage that begins with Princess Alice.

blood samples as part of medical examinations for suspected hemophilia. In addition, a Charlottesville hospital had preserved biopsy material from Anderson, taken as part of examinations for suspected cancer.

In 1994, after much legal wrangling, experts in Germany and the United Kingdom tested these samples and demonstrated that Anderson was not Anastasia. Confirmation of this conclusion came from tests performed on a lock of Anderson's hair discovered while a local historian was examining Anderson's estate.

Despite three independent test results, some supporters of Anderson's claim remain unconvinced. In any case, the mystery of Anastasia continues.

The details of the Anderson story appear in "Anastasia and the tools of justice," an editorial in *Nature Genetics*, Vol. 8, No. 3, November 1994.

Conclude this part of the activity by emphasizing that we can use DNA to establish relationships between individuals of the same species.

Homework

Complete the exercise described on the worksheet that your teacher provides.

Provide each student with a copy of Worksheet 15, *Matching DNA Sequences from Frog, Chicken, Goat, Cow, and Chimpanzee*. Instruct the students simply to follow the directions on the worksheet; the task should take no more than 20 minutes.

Part B
DNA Sequences and Life's History

Review the homework assignment by asking the students to share how they ordered their sequences. (These are actual sequence data for a 40-base segment of the beta hemoglobin gene, which is 223 bases long.) The specific order the students generated may differ, but the three mammalian sequences should be grouped together. The groupings should look something like those shown in Table T-2.

Now, ask the students to share their data on the numbers of differences between the species. You may wish to use BLM T-8, *Numbers of Differences in DNA Sequences*, to enter the data and display them as the students provide the information. Table T-3 lists the numbers of differences that the students should have discovered.

Ask: "Do you see any natural groupings to the data that you summarized in the table of sequence differences? What are they?"

The students should respond that goat, cow, and chimpanzee cluster on the basis of fewest differences in DNA sequences. Chicken and frog stand apart from this cluster.

Inform students that biologists use DNA sequences to infer evolutionary relationships between species. The evolutionary history of a species is written in its DNA, and, in the words of Niles Eldredge, "all life shares a single, complex history."[5]

Ask: "Based on the DNA data, which species are most closely related? That is, which have the fewest differences?" (goat and cow)

Next, provide each student with a copy of the beta hemoglobin sequence for humans (Worksheet 16, *DNA Sequence Data for the Beta Hemoglobin Gene in Humans*). Ask them to place this sequence in the proper place on the homework sheet. There are *no differences* in this sequence between chimpanzee and human.

You can add the human data in the space

chimpanzee	GCTGCACTGT GACAAGCTGC ACGTGGATCC TGAGAACTTC
cow	GCTGCACTGT GATAAGCTGC ACGTGGATCC TGAGAACTTC
goat	GCTGCACTGT GATAAGCTGC ACGTGGATCC TGAGAACTTC
chicken	ACTGCATTGT GACAAGCTGC ATGTGGACCC CGAGAACTTC
frog	GAAGCACGCT GAGGAACTCC ACGTGGACCC TGAAAACTTC

Table T-2 Possible grouping of DNA sequences from five different species.

	Frog	Chicken	Goat	Cow	Chimp	Human*
Frog	-	13	10	9	9	9
Chicken	13	-	6	6	5	5
Goat	10	6	-	0	1	1
Cow	9	6	0	-	1	1
Chimp	9	5	1	1	-	0
Human*	9	5	1	1	0	-

Table T-3 Numbers of differences in DNA sequences. *Students will not have the human data until you distribute Worksheet 16.

provided on BLM T-8 should you wish. Again, ask: "Based on the DNA data, which species are most closely related?" (chimpanzee and human; goat and cow)

Across the total length of the gene, these species have the following percent identity to *Homo sapiens*:

chimpanzee:	223/223	=	100%
cow:	189/223	=	84%
goat:	189/223	=	84%
chicken:	170/223	=	76%
frog:	137/223	=	61%

Use BLM T-9, *A Simple Evolutionary Tree*, to introduce students to the notion of phylogenetic trees. Emphasize that such trees propose degrees of relatedness *and* a sequence of evolutionary divergence.

Ask the students to use Worksheet 17, *Constructing an Evolutionary Tree from DNA Sequence Data*, to construct a tree for frog, chicken, goat, cow, chimpanzee, and human, based on the sequence data they have analyzed. The completed tree should look similar to that in Figure T-9.

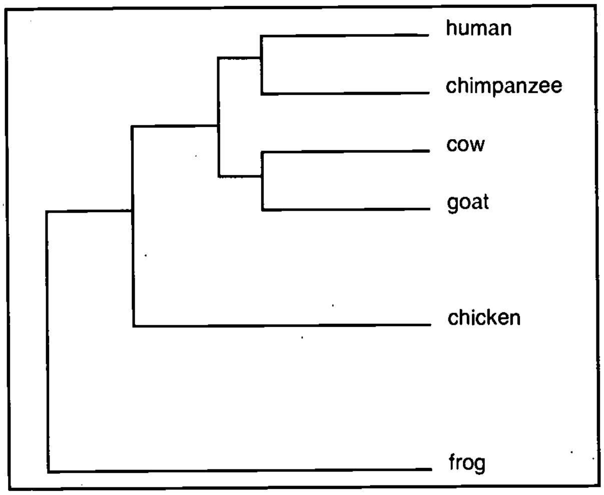


Figure T-9 Evolutionary tree for frog, chicken, goat, cow, chimpanzee, and human.

Inform the students that next they will use the computer to help them construct a tree for six different species.

PROCEDURE

Biologists use other types of sequence data to study evolutionary relationships. The next part of this activity, for example, asks you to work with amino acid sequences.

Remind the students of the relationships among DNA, amino acid sequences, and protein.

Amino acid sequences and DNA sequences are like molecular time machines that allow biologists to get a glimpse of life's history on earth. Amino acid sequences, however, allow biologists to "look back" further in time than do DNA sequences. Propose an explanation for this fact.

HINT: How many codons are there? How many amino acids are there?

This is a subtle point. Comparing either DNA or protein sequences from various organisms allows one to look back to different levels of evolutionary history. Because DNA is made up of only four bases and because the genetic code is redundant, comparisons between DNA sequences do not allow one to determine very distant relationships between species. Consider, for example, the DNA sequences: TCT AAA AGG TTA and AGC AAG CGT CTG. These two sequences show only 33 percent sequence identity, yet they both encode exactly the same amino acid sequence: Ser - Lys - Arg - Leu.[6] The comparison of DNA sequences is useful in examining relationships between individuals of the same or very closely related species in which one would expect to see little variation at the level of protein sequence. The comparison of amino acid sequences, on the other hand, allows one to examine very distant evolutionary relationships and provides some clues about protein function. If two proteins have similar biological functions but have been separated for a long time, we would expect natural selection to preserve those parts of the sequence that are directly involved in the function of the protein, while allowing changes to take place in other, nonfunctional portions of the protein sequence.

1. Work with your team members to retrieve a set of amino acid data from the computer. Follow the instructions below to bring up the screen that reads *Amino Acid Sequences for Cytochrome Oxidase.*

The password for *HGP Data and Evolutionary Biology, Part B* is <Lederberg>.

To Retrieve the Data Required for Part B

- Click on *ACTIVITY* at the upper left of your screen and highlight the line *HGP Data and Evolutionary Biology, Part B.* Release the mouse button.
- Enter the password your teacher gives you and click on *OK.*

2. This screen shows the amino acid sequences for a portion of the protein cytochrome oxidase for six different species.

Table 3 Key to abbreviations for amino acids.

A = alanine	I = isoleucine	R = arginine
C = cysteine	K = lysine	S = serine
D = aspartic acid	L = leucine	T = threonine
E = glutamic acid	M = methionine	V = valine
F = phenylalanine	N = asparagine	W = tryptophane
G = glycine	P = proline	Y = tyrosine
H = histidine	Q = glutamine	

Note that each amino acid has a one-letter abbreviation. It is not critical that you know which letter represents which amino acid, but that information is included in Table 3.

3. Note that the data are unidentified at this point. Again work with your team to replace identical amino acids with dots and to move the sequences so that the most identical sequences are grouped together.

Remind students that they are trying to maximize the similarities in the data, as they did in the activity on the Romanov family. They may notice that in this case there is no consensus sequence provided. Instead, the *Replace with Dots* command will cause the computer to replace a letter with a dot only if all of the samples have the same amino acid at that position.

4. Follow the steps below to construct a hypothetical evolutionary tree.

a. First, focus on major groups. Do the sequences cluster into a few major groups of similar sequences? How many groups do you see? Which samples belong to each group?

Students should consider the whole sequence as they identify the clusters (that is, they should not focus exclusively on the portions of the sequences that include dots). The major groupings should fall out as follows:

samples	2, 4, 6
samples	1, 3
sample	5

b. Note that the computer has not identified the species from which the sequence data were derived. Follow the instructions below to retrieve the list of species.

To Retrieve the List of Species

- Click and pull on the downward-pointing triangle that appears in the box at the top of the screen.
- Highlight the line that reads *List Species* and release the mouse button.

c. **Based on your knowledge of the species and the data from the computer, work as a team to construct a proposed evolutionary tree. Use Worksheet 18, *Constructing an Evolutionary Tree from Amino Acid Sequence Data,* to help you.**

5. **When your team has completed the worksheet, evaluate the clusters by clicking and pulling on the downward-pointing triangle that appears in the box at the top of the screen. Highlight the line that reads *Evaluate Clusters* and release the mouse button. The computer will place the species' common names next to their associated sequences. Revise your tree, if you think this new information justifies such a revision.**

The students should have developed a tree similar to that depicted in Figure T-10. Remind the students that evolutionary trees and classifications are working hypotheses about the history of life on earth and the relationship between species and higher taxonomic groups.

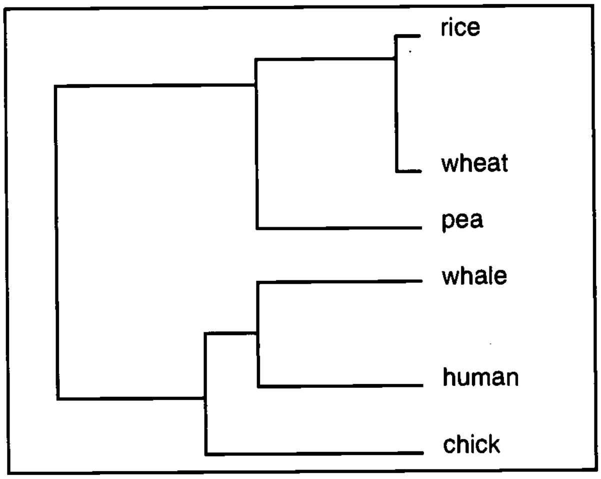


Figure T-10 Possible evolutionary tree.

Conclude the activity by emphasizing the role of the HGP in exploring evolutionary biology, and the role of informatics and computer technology in this process. In particular, emphasize that these data and the trees are only simple examples; real data are much more complex. Trees that include more species—and more sequence data—present much greater challenges in terms of databases and computational power.

NOTES:

4. The sequence data and the quoted passage are taken from Gill, P., et al. (1994). Identification of the remains of the Romanov family by DNA Analysis. *Nature Genetics* 6:130–135, used with permission.
5. Eldredge, N. (1985). *Time frames: The evolution of punctuated equilibria.* Princeton, NJ: Princeton University Press.
6. Note that the DNA sequences used in this example correspond directly to the amino acid sequence. In other words, the amino acid sequence is found by simply translating the DNA (after substituting Us for Ts) rather than by first transcribing and then translating the sequence. Molecular biologists commonly use this shorthand style when presenting coding sequences of DNA.

BLM T-1: *Target Sequences*

Target Sequence 1:

AGG

Target Sequence 2:

AGGGGA

Target Sequence 3:

AGGGGAGTCAGGGGCTCTGCATGAGGAGGG

Activity 1 Genetic Registries

BLM T-2: *Assigning Fictitious Identities*

Use this worksheet to assign a fictitious identity to each of your students. Complete one worksheet for each class of students that will be completing Activity 1. Note the following guidelines for assigning identities:

- *All of the students in any one group should be members of the same extended family.* If possible, assign at least one group to each of the families. If you have more than three groups, you can assign more than one group to one or more of the extended families. We provide space below to assign students to as many as nine separate groups, three groups per fictitious family.
- To ensure that the students in each group collect enough information to complete their pedigree, *be sure to assign at least one student to each fictitious name listed for that family.* The only exceptions to this are individuals listed as "optional." These individuals need not be assigned.
- If you have fewer students in a group than the number of names listed, you can assign some students more than one family member.

Mota/Raynes/Chen/McCarthy Family:

	Group 1	Group 4	Group 7
Sandi Raynes			
Anna McCarthy			
Kay Raynes			
Christina Chen			
Roberto Chen (optional)			

Jacobs/Schmidt Family:

	Group 2	Group 5	Group 8
Peggy Schmidt			
Katherine Schmidt			
Aisha Jacobs			
George Jacobs			
Nathaniel Jacobs (optional)			

Thomas/Major/Wray Family:

	Group 3	Group 6	Group 9
Laura Major			
Joe Major			
Cal Thomas			
Mary Jo Wray			
Bob Thomas			

BLM T-3: *Activating the Appropriate Data in the Software*

1. Click and pull on *ACTIVITY* at the top of the screen.

2. Highlight the name of the activity to which you are changing and release the mouse.

3. Enter the required password. The passwords are

   The HGP and Electronic Databases ______________________________

   Genetic Registries ____

   Explaining the Outliers, Part A ____________________________________

   Explaining the Outliers, Part B ____________________________________

   Genetic Anticipation ____

   HGP and Evolutionary Biology, Part A ________________________

   HGP and Evolutionary Biology, Part B ________________________

4. Click on *OK* and the computer will activate the data for the new activity.

Dear Clients:

We have conducted DNA analysis on the samples you provided. Our results indicate that

1 person in the test group carries the dF508 mutation for cystic fibrosis. This person is heterozygous; there are no clinical implications for this person.

1 person carries the T ➡ A (glu ➡ val) mutation for sickle hemoglobin. This person is heterozygous; there are no clinical implications for this person.

1 person carries the allele for cardiomyopathy. This person is at substantial risk for sudden heart attack. He or she should see a physician immediately for counseling about health care and prevention. It is likely that this person should restrict his or her physical activity severely.

Option about the Disclosure of Anonymous, Combined Data

Option 1: Report only anonymous, combined data about the total number of people in the group who carry the alleles for CF, SCD, or CM.

reasons to make public	reasons not to make public

Options about the Disclosure of Individual Test Results

Option 2: Report each set of individual results only to the person tested.

reasons to release

reasons not to release

Option 3: Announce the individual results openly or publicly to all.

reasons to announce

reasons not to announce

Option about Entering Individual Test Results into the LGD:

Option 4: Enter the individual results into the LGD.

reasons to enter the data	reasons not to enter the data

BLM T-6: *Instructions for Interpreting the Results of DNA Testing*

General Instructions*

Examine the envelope carefully before you open it to determine that it is labeled with your correct name and signature. **If you have any questions about the envelope or if you are not sure that the envelope belongs to you, return it to your teacher immediately**.

Open the envelope carefully and look inside to determine the color of the tips of each swab. Do not remove the swabs from the envelope. Use the key below to determine the results of each test.

Color	Interpretation
White	White tips indicate that the testing did not reveal the presence of any known disease-related allele.
Red	Red tips indicate that the testing revealed the presence of at least one allele known to be associated with cystic fibrosis. Individuals who receive a red-tipped swab should seek qualified genetic counseling to discuss the significance of this result.

Blue Blue tips indicate that the testing revealed the presence of at least one allele known to be associated with sickle cell disease. Individuals who receive a blue-tipped swab should seek qualified genetic counseling to discuss the significance of this result.

Green Green tips indicate that the testing revealed the presence of at least one allele thought to be associated with familial hypertrophic cardiomyopathy. Individuals who receive a green-tipped swab should seek qualified medical attention and genetic counseling immediately to discuss the significance of this result.

*These instructions are provided as a service of GeneTest, Inc. and are not to be used in the interpretation of any results except those provided by GeneTest, Inc. Individuals receiving these results should consult a qualified health-care professional for help in interpreting test results. GeneTest, Inc. is not liable for any damages resulting from improper interpretation of test results.

Recommendation A:

Goal of the recommendation:

Works well in the PKU case? (circle one) yes no
Reason:

Recommendation B:

Goal of the recommendation:

Works well in the PKU case? (circle one) yes no
Reason:

Recommendation C:

Goal of the recommendation:

Works well in the PKU case? (circle one) yes no
Reason:

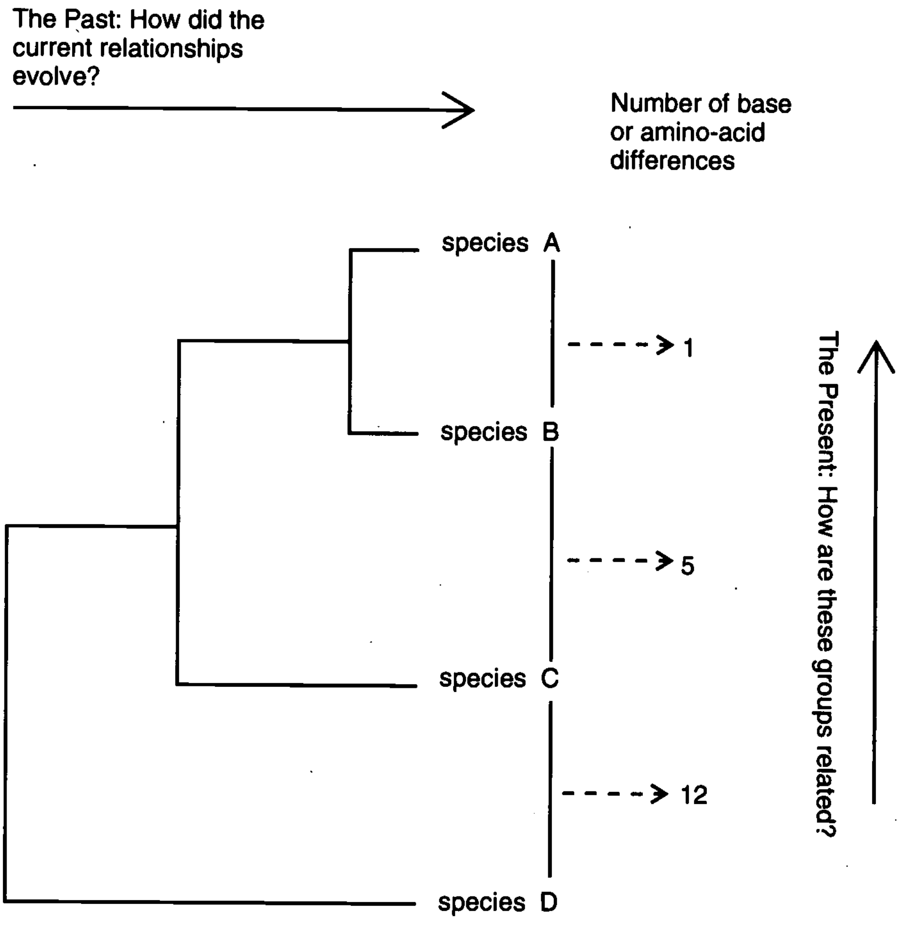
Extension Activity HGP Data and Evolutionary Biology

BLM T-8: *Numbers of Differences in DNA Sequences*

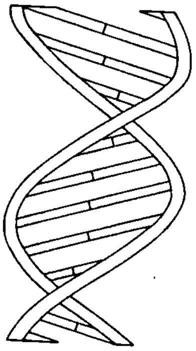
	Frog	Chicken	Goat	Cow	Chimp	Human
Frog						
Chicken						
Goat						
Cow						
Chimp						
Human						

Extension Activity HGP Data and Evolutionary Biology

BLM T-9: *A Simple Evolutionary Tree*


The Past: How did the current relationships evolve?
Number of base or amino-acid differences
species A
species B
species C
species D
1
5
12
The Present: How are these groups related?

An evolutionary tree indicates the degree of relatedness among species. It also provides an historical picture of the likely sequence in which the species diverged over time. The tree, therefore, provides a sense of current phylogenetic relationships and a sense of how those relationships came to be.



Introductory Activity The HGP and Electronic Databases

Imagine that you are a member of a team of scientists that is searching for a particular gene. You have just sequenced a short section of DNA that you think might be a piece of the gene, and now you need to search the published sequence data to see whether anyone else already has reported the same sequence.

Procedure

1. Follow your teacher's instructions for completing a manual search (a search by hand) of the sequence data that you have been given.
2. Follow your teacher's instructions to bring up the sequence data stored in your computer. Complete the steps listed in the box below to repeat the searches you just conducted manually, this time having the computer do the searching.
   a. Search the sequence data for the first target sequence (AGG). What data does the computer retrieve for you?

To Search the Sequence Data for a Target Sequence

- Click and pull on the box labeled *TARGET SEQUENCE* to highlight the sequence of interest. (**Windows Users**: Click on the arrow.)
- Release the mouse button.
- Click on *BEGIN SEARCH.*

NOTE: Click on the scroll bar arrow at the right of the screen to see the whole sequence.

   b. Click on *COUNT* to determine the total number of AGG sequences that the computer found. Record this number and compare it with the number that you found during your manual search.
   c. Now search for the slightly longer sequence, AGGGGA. Again, click on *COUNT* to determine the number of matching sequences that the computer found. Record this number and compare it with the number of matches that you found during your manual search.
   d. Finally, select the full sequence of interest: AGGGGAGTCAGGGGCTCTGCATGAG-GAGGG. How many times does this sequence occur in your data? How long did it take the computer to locate the sequence?

Questions for Discussion

1. How does the difficulty of the search change as your target sequence gets longer?
2. What do the results of the two different types of searches suggest about the advantages of computer-based storage and retrieval of sequence data?

Homework Assignment

Read the background information for Activity 1 (Worksheet 3, *Genomic Databases*) and answer the study question provided.

BLM: Worksheet 1
Introduction to the Human Genome Project

The Human Genome Project (HGP) is the first large coordinated research effort in the history of biology. Currently involving thousands of scientists in the United States and around the world, its primary objective is to create a detailed *map* of the human genome. This genome contains the hereditary instructions that guide the development of a fertilized egg into a complex organism of more than 10^{13} cells.

For the purpose of the HGP, **genome** means one each of the different types of chromosomes found in most human cells: 22 different autosomes, plus the X and Y chromosomes, plus the mitochondrial chromosome. Geneticists estimate that this genetic material includes between 50,000 and 80,000 genes. To map the genome completely, scientists will have to identify each of these genes and determine its precise location on a particular chromosome.

Another objective of the HGP is to determine the complete base sequence of the human genome. This means determining the exact order of the approximately three billion As, Ts, Cs, and Gs that make up the DNA of the different types of human chromosomes, as well as cataloging the major base sequence differences that occur from one of us to the next. This is an enormous task. By early 1996, geneticists had identified only about 6,000 human genes, and had determined the base sequence(s) of only about 2,000 of these.

Scientists expect that new data generated in conjunction with the HGP will help them address a number of interesting biological questions. An obvious benefit of the work is the discovery of new genes. As laboratory groups identify and sequence these genes, scientists build a foundation for studying gene function and regulation and, possibly, for discovering and investigating new biological processes. Access to detailed map and sequence data also will affect the practice of clinical medicine. For example, finding the genes associated with inherited diseases is a key concern of the HGP. The discovery of these genes likely will lead to the development of rapid, inexpensive techniques to diagnose genetic disorders in individuals who exhibit certain symptoms and also to the development of a variety of DNA-based tests that will help geneticists detect carriers. Our growing knowledge of human genes and gene products also is giving the pharmaceutical industry powerful new tools to help physicians treat many inherited disorders. Although isolating and sequencing a disease-related gene does not guarantee that scientists will be able to develop a treatment, it can help speed the process significantly.

As work on mapping the human genome moves forward, scientists involved in the HGP also are studying the genomes of other species including a bacterium, a yeast, a nematode (roundworm), a fruit fly, a mouse, and a plant. Comparing these genomes with the human genome is helping scientists find human genes more easily, is deepening our understanding of how genes function, and is providing important new information about molecular evolution, including human evolution.

Study Questions

1. When biologists talk about the human genome, they often use terms such as **gene**, **chromosome**, and **base** to refer to different levels of organization of the genetic material. Using each of these terms correctly, describe what scientists must accomplish in order to map and sequence the human genome. Be sure that your answer illustrates clearly the relationships that exist among these levels of organization.

2. You probably have learned that most human cells contain 46 chromosomes (notable exceptions include eggs and sperm). You also probably have learned that different forms of the same gene—for example, different **alleles** of the gene for cystic fibrosis—contain different sequences of bases.

   a. Explain how it is, then, that scientists studying the human genome will map only 25 different types of chromosomes.

b. What do you think scientists studying the human genome mean when they talk about sequencing "the" human genome? Does "the" human sequence actually exist?

3. Although the focus of the HGP is on mapping and sequencing the human genome, scientists involved in the effort also are studying the genomes of several other organisms.

a. Table 1 lists the approximate sizes of the genomes of several of the HGP organisms. If you assume an average sequencing rate of 10 million bases each year, how long would it take to sequence each of these genomes?

b. The background information we provide identifies some of the benefits that scientists are realizing by mapping and sequencing these nonhuman genomes. What do these benefits tell us about the organization and function of these nonhuman genomes in comparison with the organization and function of the human genome? What do they tell us about evolution?

Table 1 Approximate sizes of the genomes of several HGP organisms.

Type of Organism	Estimated Numbers of Base Pairs (millions of pairs)
bacterium	4.7
yeast	15
fruit fly	80
human	3,000

BLM: Worksheet 2
Finding a Sequence

Read the sequence from left to right across the page.

```
CTACGGTGACAGCTGCCAGGATCCTAAAAGGGCAGAAGAAGGACAAACTGGGGCCTGAGACCTTAG
GGGCCATGGACCGCTTCCCGTACGTGGCTCTGTCCAAGGTAAGTGCTGGGCTACCTTAGAGTCCTC
CAAGCAGAGAAGGGGAATCCTGGCTATGGAGTGTGGTAGGAGGGAGGGACCCTAAACAGCTGGGGC
TCCAATAAGGAGCTGGAGGCAGTTGGAATCCCAGAGGACAGAGATCAGGGTCTTGTTTGTCTGCCC
CAGAGAAGAGCTCAGAGTGTCTCTGTCCCCAGACATACAGTGTAGACAAGCATGTGCCAGACAGTG
GAGCCACAGCCACGGCCTACCTGTGCGGGGTCAAGGGCAACTTCCAGACCATTGGCTTGAGTGCAG
CCGCCCGCTTTAACCAGTGCAACACGACACGCGGCAACGAGGTCATCTCCGTGGTGAATCGGGCCA
AGAAAGCAGGTGGAGCTGGGGCCCGGCTGTGGGGTCAGGGCCAGTGACAGACCTCTATCGCATATC
CTGACCTCTATCACCCTCAGGAAAGTCAGTGGGAGTGGTAACCACCACACGGGTGCAGCATGCCTC
GCCAGCCGGCACCTACGCCCACACGGTGAACCGCAACTGGTACTCGGATGCCGACGTGCCTGCCTC
GGCCCGCCAGGAGGGGTGCCAGGACATCGCCACGCAGCTCATCTCCAACATGGACATTGATGTGCG
ACCCCCGGGCCAAGGGTGGGGCTGGGCAGAGAGTAGCAGGGAGGGGGCACCAGCTCAGACCAGGCA
ACCAAAAGCCTTATCTGGGCCAGCAGGGTCTGGAAGGTGGGGTTGGGGGCGTAGAAGGCGCACCAG
GCTGGGCCATTCCCACAGCCTTGGGGAGGGGAGTCAGGGGCTCTGCATGAGGAGGTGACACGGGGC
CTAGCCATGGCCCAAAGTCCACCTGCCCCATCCTCTGTTCCCAGGTGATCCTAGGTGGAGGCCGAA
AGTACATGTTTCCCATGGGGACCCCAGACCCTGAGTACCCAGATGACTACAGCCAAGGTGGACCA
GGCTGGACGGGAAGAATCTGGTGCAGGAATGGCTGGCGAAGCACCAGGTGATGGGGGCTGGTGGGT
GTGCTGGGCACAGCAGGGGGAGGGCAGAGGTGTGGGGCTCGGGGCTGTGGGCTGAGGCCTGGCTCT
CTCCCTCCCCGCAGGGTGCCCGGTACGTGTGGAACCGCACTGAGCTCCTGCAGGCTTCCCTGGACC
CGTCTGTGACCCATCTCATGGGTAATGACCCCCTTCCTGCCCTGGCATCCTCAGATGGCCTCAGAT
GGCACTTCTGAGCCTGTGTGCACATCCGCCAGCACCCTCCCACCCCCAGCCTGCCAGTCACCACAG
GACCCCTTGTCCCACAGGTCTCTTTGAGCCTGGAGACATGAAATACGAGATCCACCGAGACTCCAC
ACTGGACCCCTCCCTGATGGAGATGACAGAGGCTGCCCTGCTCCTGCTGAGCAGGAACCCCCGCGG
CTTCTTCCTCTTCGTGGAGGGTGCGTGGTGGCCCTGGGAGTGGGGGGTTGGGGGTTGGAGCAGGGC
AGGCTCAGCATCTCCCCCCTCTGGCCTTCCTGCAGGTGGTCGCATCGACCATGGTCATCATGAAAG
CAGGGCTTACCGGGCACTGACTGAGACGATCATGTTCGACGACGCCATTGAGAGGGCGGGCCAGCT
CACCAGCGAGGAGGACACGCTGAGCCTCGTCACTGCCGACCACTCCCACGTCTTCTCCTTCGGAGG
CTACCCCCTGCGAGGGAGCTCCATCTTCGGTAGGCCTGGGGATGAGTGGCAGGTGCTGCTGCAGCA
ATTAAGTGGGTGAAATCTGAGCCTCAGTCTCCTCCTCTGTCAAGTGGGAGTAATGCTGGCACCAGC
CTAATAGGGTCCTCTGCGGACTAAGCCCCTGACCAGGCAAAACGTGCGGTGCCTAGCACGTGGGAG
ACACTCCACAGCTGTGTTCAGCTCAACCACAGGGACCCCTCTCTCAGGGGAGTCAGGGGCTCTGCA
TGAGGAGGGCAGGAAGGCCTACACGGTCCTCCTATACGGAAACGGTCCAGGCTATGTGCTCAAGGA
CGGCGCCCGGCCGGATGTTACGGAGAGCGAGAGCGGTGAGTGCCGTGGGGTGGCCTGAGGGGGACC
AGGGTGCCAAGGATGGGGGGCTGGCGGGAAGGGGTCACCTCTTGTCTGCCTGGAACTGAAACTTCC
TACTGAAACTGAACCCTCCAACCAGGGAGCCCCGAGTATCGGCAGCAGTCAGCAGTGCCCCTGGAC
GGAGAGACCCACGCAGGCGAGGACGTGGCGGTGTTCGCGCGCGGCCCGCAGGCGCACCTGGTTCAC
GGCGTGCAGGAGCAGACCTTCATAGCGCACGTCATGGCCTTCGCCGCCTGCCTGGAGCCCTACACC
GCCTGCGACCTGGCGCCCGGCACTTCTGAGCCTGTGTGCACATCCGCCAGCACCCTCCCACCCCCA
```

BLM: Worksheet 3
Genomic Databases

As difficult as the tasks of mapping and sequencing the human genome are, a more intimidating problem is what to do with the huge amount of information that will result from all of this research. Just to print one letter to represent each of the three billion bases in the human genome would take the equivalent of 200 telephone books of 1,000 pages each.

How can scientists store such a massive amount of information? How can they analyze it and come to understand what all of it means? And how can these data be made available to the wider scientific and medical communities? Workers involved in the HGP already have made many new and useful discoveries, including identifying and sequencing the genes for Huntington disease and myotonic dystrophy. How can new information be communicated rapidly to those who can benefit from it?

A group of scientists and information specialists provided a partial answer to these questions when they introduced the Genome Database (GDB) to participants at a workshop on human gene mapping in Oxford, England, in September 1990. GDB is a computerized database that organizes, stores, and allows the retrieval of gene mapping data. You might think of it as a comprehensive, electronic dictionary that serves as the official repository for all of the mapping data generated by the HGP. By early 1996, GDB, which is maintained at Johns Hopkins University in Baltimore, Maryland, contained information on more than 5,900 human genes. As the volume of map data increases, GDB's importance as a facility for storing and sharing these data also increases.

A scientist can use GDB in a variety of ways. If she is working on the HGP, she may enter new data into the database as they are generated. (Because GDB is a public database, all of us may access it for reading purposes. Only specially designated individuals, however, may enter or change information. This helps prevent the entry of incorrect data.) As a geneticist, physician, or genetic counselor, she also may draw upon the information already stored in the database to help her in her own research or to help her find answers to genetic questions involving patients. Used properly, GDB can help scientists quickly learn about discoveries made in other laboratories and also can help them avoid duplicating the efforts of other investigators who may be interested in the same research questions or problems.

Another part of the answer to the problem of storing and communicating huge amounts of genomic information was provided with the creation of GenBank, another research database important to the HGP. Scientists and information specialists developed GenBank in the early 1980s, years before the HGP began, to store DNA sequence data that had been derived from all types of organisms, from bacteria to humans. Since the start of the HGP, however, GenBank has become a key repository and source of sequence data on human genes.

Scientists search the data in GenBank, which is now located at the National Center for Biotechnology Information in Bethesda, Maryland, to determine whether the sequence of a newly identified piece of DNA matches that of any previously reported pieces. This type of search sometimes can help scientists identify new genes or new alleles of genes or even help them determine the function of a specific piece of DNA. By early 1996, scientists had entered into GenBank information about the base sequences of more than 680,000 individual pieces of DNA. Together, the total number of bases in these pieces was more than 460 million.

GDB, GenBank, and other **research databases** store data that are anonymous. This means that although we know that some of the data describe human genes, we cannot connect these human data to any one specific person. That is, we do not know the person from whom each piece of sequenced human DNA was derived.

A second type of database that is used in the HGP, however, is a **registry database**, a database in which the stored information is tied to specific individuals. You may be familiar already with a number of different types of registries including computerized mailing lists that link information about your buying habits to your name and address, and state driver's license databases that allow police immediate access to information about your driving record.

Many scientists and health-care professionals use registry databases to store, organize, and analyze genetic and health-related information on specific individuals and groups of individuals. A genetic counselor, for example, may use a database containing information he has collected about his own clients during a counseling session. Likewise, investigators working in conjunction with the HGP often create large registries to store and track genetic information about individuals who have specific genetic diseases. For example, scientists in France recently used a registry database to identify the inheritance pattern of a gene for hereditary juvenile glaucoma. Sometimes, the first clue that a specific segment of DNA contains a disease-related gene comes from an extensive analysis of this type of personal genetic data.

Because the information in a registry often is sensitive, access to registries usually is restricted to specific people who need the data and who have been given special permission to handle it. The armed forces of the United States store DNA from each of their members and limit access to this DNA to those who may need the information to identify a specific individual. This DNA registry helped authorized personnel identify many of the Americans killed during the 1991 Persian Gulf War. Usually, the creators of a registry list the conditions under which data can be released before they ask people to sign consent forms allowing the collection and storage of this information.

As you complete the remaining activities in this module, you will work with the following two model databases: first, the Local Genome Database (LGD), which represents a registry and, second, the National Genome Database (NGD), which represents a public, nationally supported research database. Because the LGD and the NGD were developed specifically for this module, their features do not match exactly any of the databases actually associated with the HGP. Instead, the LGD and the NGD are model databases with structures and content that illustrate the actual databases that are used in the HGP.

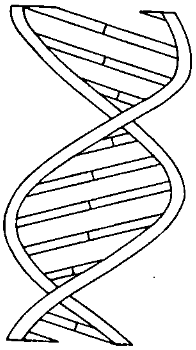
Study Question

Use the information provided in the reading to develop your own definitions of the following terms:

a. database

b. research database

c. registry database



Activity 1
Genetic Registries

Imagine that you live sometime in the future, in a small town somewhere in rural America. Like other small towns (as well as big cities), your town finds itself increasingly affected by the rapid increase in genetic information that has resulted from the HGP. Members of the community, for example, can access a wealth of genetic data simply by searching the National Genome Database (NGD), a research database that is available online to anyone around the world who has the basic computer and telecommunications capabilities required to access it.

Imagine as well that about a year ago the public-health authorities decided to collect the growing body of personal, medical, and genetic data they had accumulated about the people of your community into a central database called the Local Genome Database (LGD). Doctors, nurses, and public-health officials can access this database to retrieve information important to their work.

Concerned about the public's response to this action, health authorities decided to make a small subset of these community health data available online. Their reasoning was that this would allow community members to learn about this resource and also give interested individuals a chance to explore the database and to provide input about the appropriate uses of these data. In the interest of community education, three families in the town volunteered to allow portions of their records to be made available to the public. These records include each family member's genotype for the following four genes: the fragile X gene (F); the gene for hereditary juvenile glaucoma (G); the angiotensinogen gene, a gene that can increase significantly an individual's risk for high blood pressure (A); and the gene for alpha hemoglobin (H).

Procedure

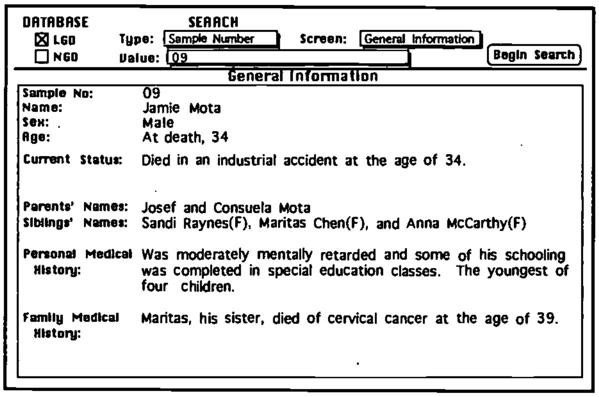
1. Your teacher will give you the name of the fictitious person whose identity you will assume for this activity. Your tasks are
   - □ to locate and examine your personal file in the LGD,
   - □ to accumulate sufficient information to develop a pedigree that properly represents the structure of your extended family (your parents, siblings, aunts, uncles, cousins, and grandparents),
   - □ to begin to interpret the genetic data you find in the LGD, and
   - □ at the end of the activity, to decide whether you would like your fictitious person to be tested for his or her genetic profile for three additional genes.

2. Take turns following the steps below to find the information that you need to complete Worksheet 4, *Collecting Family Data*.

   a. Locate your "personal file" in the LGD.

   b. The screen that appears should be similar to the screen pictured in Figure 1. What types of information does it give you?

   c. Write your sample number and your fictitious name, age, parents' names, and sibling's or siblings' name(s) and sex(es) into the appropriate areas on your worksheet.



Sample No: 09
Name: Jamie Mota
Sex: Male
Age: At death, 34

Current Status: Died in an industrial accident at the age of 34.

Parents' Names: Josef and Consuela Mota
Siblings' Names: Sandi Raynes(F), Maritas Chen(F), and Anna McCarthy(F)

Personal Medical History: Was moderately mentally retarded and some of his schooling was completed in special education classes. The youngest of four children.

Family Medical History: Maritas, his sister, died of cervical cancer at the age of 39.

Figure 1 Sample *General Information* screen from the LGD.

NOTE: Although each of you should complete your own computer search, you should work together to be sure that each member of your group retrieves the correct information. You will need this information to complete your family's pedigree.

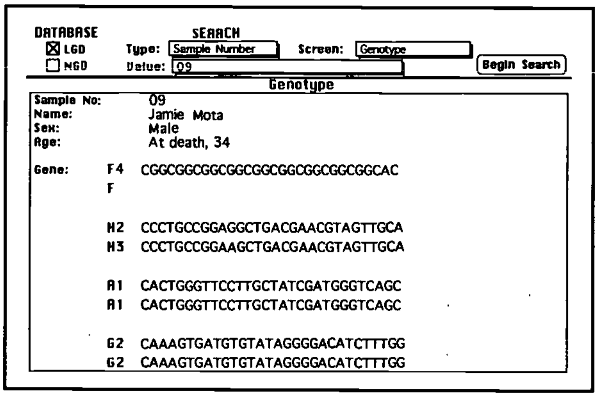
3. Use the data that you collected on Worksheet 4 to construct a pedigree for your extended family.

NOTE: You will find instructions for building your pedigree on Worksheet 5, *Constructing a Pedigree*. You will need the pedigree that you develop to complete Activity 2.

4. Take turns following the steps below to examine your personal LGD file more closely.

   a. Again, locate your personal file in the LGD.

   b. Change to the *Genotype* screen to see your DNA profile for four different genes. The screen that appears should be similar to the screen pictured in Figure 2. What types of information does it give you?



Sample No: 09
Name: Jamie Mota
Sex: Male
Age: At death, 34

```
Gene:   F4   CGGCGGCGGCGGCGGCGGCGGCGGCGGCAC
        F

        H2   CCCTGCCGGAGGCTGACGAACGTAGTTGCA
        H3   CCCTGCCGGAAGCTGACGAACGTAGTTGCA

        A1   CACTGGGTTCCTTGCTATCGATGGGTCAGC
        A1   CACTGGGTTCCTTGCTATCGATGGGTCAGC

        G2   CAAAGTGATGTGTATAGGGGACATCTTTGG
        G2   CAAAGTGATGTGTATAGGGGACATCTTTGG
```

Figure 2 Sample *Genotype* screen from the LGD.

   c. Discuss the information you that find on the *Genotype* screen and answer the questions on Worksheet 6, *Interpreting Data in the LGD.*

To Locate a Person's File in the LGD

- Click on the *LGD* checkbox.
- Click and pull on the *TYPE* box to highlight the line that reads *Name*. Release the mouse button.
- Click on the *VALUE* box. The computer will display a box that asks you to enter the name of the person for whom you wish to search.
- Enter your fictitious name (first and last) in the box and click on *OK* (*BEGIN SEARCH* in Windows). If there already is a name in the box, simply type over it.

NOTE: Remember that the computer searches for exact matches. Spaces count. If you make an error, press the [delete] key located on the upper right side of the keyboard. This will delete the last letter you entered.

TIP: You also may enter just the last name of your fictitious person and retrieve a list of everyone in the database with that last name. To retrieve the personal file of an individual whose name appears on the list, simply click on his or her name. To return to the list, click on the *RETURN TO LIST* button that will appear in the upper right of the screen. (**Windows Users:** Double click on a person's name on the list to retrieve his or her file. Click on *BEGIN SEARCH* to return to the list.)

To Change to the Genotype Screen

- Click and pull on the *SCREEN* box to highlight the line that reads *Genotype*.
- Release the button. The computer will display the new screen.

NOTE: When you first access a person's file, the computer always will display the *General Information* screen. To move back to this screen, click and pull on the *SCREEN* box, highlight *General Information*, and release the button.

5. At the end of the period, your teacher will offer you an opportunity to have your fictitious person "tested" for his or her genotype for each of three additional genes (the genes associated with cystic fibrosis, sickle cell disease, and familial hypertrophic cardiomyopathy). Consider your decision carefully and follow your teacher's instructions for either consenting to the test or for refusing it.

Homework Assignment

Read the background information for Activity 2 (Worksheet 7, *Kate and Ryan*) and answer the study questions provided.

BLM: Worksheet 4
Collecting Family Data

Complete the table using information that you retrieve from the LGD. Record only the information you retrieve for your fictitious person. In the next step, you will share this information with the other members of your group.

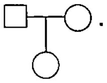
Notice that the information in the second column is only a sample of the type of information that you should retrieve. This person may or may not be a member of your fictitious family.

Sample number:	**09**	
Fictitious name:	**Jamie Mota**	
Age:	**at death, 34**	
Parents' names:	**Josef & Consuela Mota**	
Siblings' names and sexes:	**Sandi Raynes (F), Maritas Chen (F), Anna McCarthy (F)**	

BLM: Worksheet 5
Constructing a Pedigree

A family history, or pedigree, is an important tool in human genetics research. Pedigrees help us visualize the relationships that exist among various family members, and also help us trace the hereditary patterns associated with specific genetic traits. As we begin to see these hereditary patterns across many families and many generations, we may gain new insights into the nature of the genes involved.

How to draw a pedigree. Scientists use the symbols shown in Figure 3 to draw pedigrees. Notice that

- males are represented by squares, □;
- females are represented by circles, ○;
- *parents* are joined by a horizontal line drawn from the *middle* of each symbol, □—○ ; and
- *children* are connected to their parents by lines drawn to the *top* of each symbol, .

Notice also that everyone in *one generation* is shown on the same horizontal row. For example, in the pedigree shown in Figure 3, the man and woman represented in the first row (generation I) are the *parents* of the woman shown in the second row (generation II), and the *grandparents* of the individuals in the third row (generation III).

Procedure

Follow the steps below to construct a pedigree for your fictitious family. You may want to draw your pedigree in pencil so that you can change the drawing easily if you make an error.

NOTE: During your work on the computer, your team should have retrieved all of the information that you need to construct this pedigree.

1. **Draw the pedigree for your immediate family.** Follow steps a - e below to construct a small pedigree showing your immediate family. Draw this pedigree in the empty box to the right.

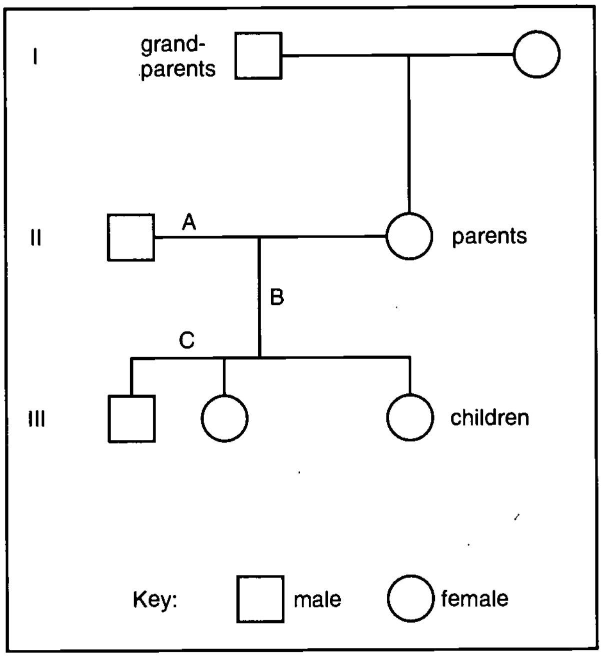


Figure 3 Sample pedigree with generations labeled.

a. Draw a circle to represent your mother and a square to represent your father. Connect these two symbols with a horizontal line (see line A in Figure 3). Label each symbol with the person's name.

b. Draw a vertical line that extends down from line A (see line B in the sample). This vertical line will connect your parents to their children.

c. Draw a second horizontal line at the end of line B (see line C in the sample). This line will connect all of the symbols for your parents' children.

d. Look at the information that you have about yourself and about your brothers and sisters. To add all of you to the pedigree, draw a row of circles and squares (as appropriate) below line C and label each with the proper name. One of these symbols should represent you, and the rest of them should accurately show your brothers and sisters.

e. Draw a series of short vertical lines from line C to the top of each symbol. These lines connect the symbols for you and your siblings to the pedigree.

2. **Draw the other family units that are represented in your group.** Now draw a pedigree for each of the other individual families represented in your group. (You will have to copy the individual pedigree from each of the other members of your group). Draw these pedigrees in the box below. At this point, do not try to connect these families. Label each symbol with the person's sample number and name.

3. **Identify the generations represented.** Follow steps a and b below to organize the pedigrees you just drew.

   a. Look carefully at all of the individual family pedigrees that you have drawn (don't forget your own pedigree that you drew in Step 1). You should find that some people appear in each of two families. Circle the names of these individuals.

   b. Notice that the people whose names you have circled are ali brothers and sisters. Their parents are the grandparents in your extended family. Place a check mark over the symbols for the grandfather and grandmother.

4. **Build a complete pedigree for your extended family.** Build a complete pedigree by connecting each of the small pedigrees you have constructed already. Begin by drawing the people that you identified as grandparents. Then add their children (the people whose names you circled and their siblings), and this time draw them together with their wives or husbands and their children. If you complete your pedigree properly, you should see three generations.

BLM: Worksheet 6
Interpreting Data in the LGD

Take turns examining the *Genotype* screens for your fictitious people. Then discuss and answer the following questions.

1. Notice that your person's record contains a list of base sequences for genes F, H, A, and G. Why does this record show two sequences for most or all of these genes?

2. The sequences shown are only 30 bases long. Do you think that each sequence represents one complete gene or only a portion of a gene? Explain your answer.

3. What explanation can you offer for the observation that males have only one F gene and never two?

4. Consider genes H, A, and G. For which of these genes are you homozygous? For which are you heterozygous? List two ways in which you can determine this from the information on the screen.

5. Notice that the LGD does not give you any information about the genes symbolized by these letters. What type of database would you have to search to find information about these genes?

BLM: Worksheet 7
Kate and Ryan

Kate Dozark received a call from Stanley Williams, a genetic counselor, asking if she and her husband Ryan could come in to have a repeat blood sample taken. Kate and Ryan had visited the genetic counselor the week before because their oldest son, Chris, has cystic fibrosis (CF) and they were considering prenatal testing as one choice they might make in a future pregnancy. The purpose of the blood test was to evaluate their suitability for prenatal testing.

When Kate and Ryan arrived at the office, Stanley could tell that Ryan was annoyed. "Is anything troubling you?" Stanley asked. "Well," said Ryan, "we were just in here a week ago for the same thing and now we're back again. Why do we have to make so many visits?"

Stanley explained that a piece of laboratory equipment that was processing their samples had broken during the test. Laboratory personnel were concerned that some of the samples may have become contaminated or mixed up, and, in reporting the results, had suggested that the geneticists should be aware that there could be a problem. Stanley told Ryan and Kate that this was a responsible action by the laboratory. "After I reviewed the results," he said, "I thought that there was a problem with the samples. I think the test needs to be repeated."

Ryan did not like having his blood drawn, was upset that Chris has CF, and had been short-tempered a lot lately. "I don't understand how you could tell from the test results that there might have been a mishap in the lab. I still want to know why you need more blood," said Ryan.

"Let me go over the information we have on you, Kate, and Chris, and I think you'll see what I mean. Here's your family's pedigree, together with the results we got back from the lab."

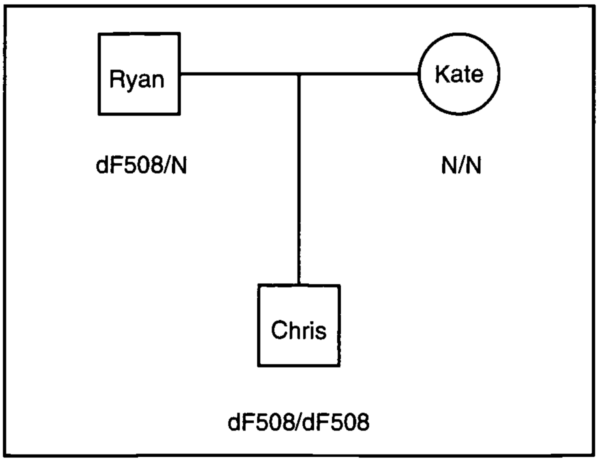


Figure 4 Pedigree and laboratory results for Kate, Chris, and Ryan.

"The dF508 mutation is the most common of the several hundred possible CF mutations," said Stanley. "You can see that Ryan is a carrier for this mutation (he has one dF508 allele) and that he also has one N, or normal, CF allele. We know from Chris's test results that Chris has CF and is homozygous for the dF508 mutation. That means that both of his parents must carry the dF508 allele, but only Ryan showed this on his test. Kate's test showed her to have two N alleles. From what you know about recessive inheritance patterns, you know this doesn't make sense. We need to look for an explanation of these unusual results, because we're sure that Kate is Chris's mother."

Ryan nodded "yes" at the memory of how proud he was of Kate at Chris's birth. Even though her labor was long and hard, she hardly complained at all. In fact, as he remembered the day, it was Kate's calm attitude that helped hold *him* together.

"The lab already has given us a hint that something may have gotten confused, and by repeating the test, we can be sure," continued Stanley. Stanley went on to explain what went into a genetic test and how, although there were a lot of places where things could go wrong, most of the time the tests went smoothly. "First, we talk with the family, just as we did with you last week, and we go over the reasons for doing a test, what we might expect to find, how long it will take, how much it will cost, and what types of things we can do with the information. If the family decides to go ahead with the test, we make sure that we have an accurate family history (a pedigree). It is important that all the family relationships be described in genetic terms. Sometimes, parents are foster parents or adoptive parents or stepparents, and thus are not genetically related to the child, and we need to know this."

"Once we have the pedigree, we take a blood, hair, or skin sample and place it into a tube labeled with the person's name and sometimes a hospital number. Someone carries the tube down to the laboratory where it goes into a basket with many other samples. Eventually a technician takes it out and records the name and number into a logbook. The sample is then processed, in one or more large machines, sometimes together with hundreds of other samples. In some cases, part of the sample is removed and placed into a new tube that is labeled with a new number or letter code. At the end of the tests, the technician views the results on a film or by a color change in the tube, and enters the data into a database. This data-entry technician may enter thousands of results a day into the computer, although new technology sometimes lets the computer interpret and record the results directly into our hospital's LGD. After this, the results are sent on a paper slip to be entered into the hospital chart, or a printout of the results is placed in the patient's records. Finally, when the results come back, I check to make sure they fit with the pedigree just as we did today. If there are any inconsistencies, we work hard with the family and the lab to resolve them. Making a mistake here can have pretty serious effects on lots of people."

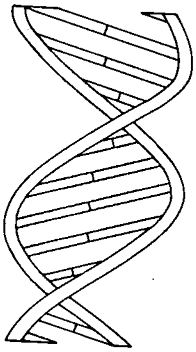
"From the time a sample leaves my office until I get the results back, ten or more people may have looked at it or helped process or record data about it. It also has traveled along with hundreds of other similar samples. Because so many people and instruments handle so many samples, we train people to be very careful in recording data and to remember always that this information is confidential. With all the things going on, we are proud of the fact that it is very seldom that something goes amiss. Unfortunately for you, that did happen, but we hope that it won't happen again."

Stanley assured Ryan that the problem in this case was nobody's fault. "We're sorry this has happened. Of course, there will be no charge for the repeat blood test." Stanley asked if they had any questions. Ryan did not have any questions, but thanked Stanley for his time and explanation and said he hadn't realized how complicated a laboratory test could be. Chuckling, he added "It also explains why tests sometimes seem to cost so much!"

A year and a half later, Chris came home from school to learn that he had a new brother Doug (dF508/N, just like his parents) and two very happy parents.

Study Questions:

1. Why did Stanley think there had been a mix-up in processing Kate's sample? Use information provided in the pedigree on p. S-15 in your answer.
2. Suppose Kate's test had shown her to have the genotype dF508/dF508. Do you think Stanley would have questioned this result? Explain.



Activity 2
Explaining the Outliers

One of the goals of the introductory activity was to give you some sense of the nature and the volume of genomic information that researchers associated with the HGP are generating. Without computers, these data already would be unmanageable.

In Activity 1, you used the information that you retrieved from the LGD to construct a pedigree for your extended family. Scientists and health-care professionals use pedigree data in many ways. For example, investigators interested in identifying relationships between genes and specific human traits (including disease conditions) may use pedigree data to help establish the heritable nature of a trait and to determine the specific pattern of inheritance involved. Likewise, genetic counselors and other health-care professionals may construct pedigrees to help explain the principles of inheritance to members of families who carry genes associated with known diseases. The possibility exists that some day public-health officials also may use pedigree data to trace individuals at risk for certain preventable genetic conditions.

In this activity, you will have the opportunity to investigate the data in the LGD more thoroughly. You also will access a research database, the National Genome Database (NGD), for information that will help you better understand the data in the Local Genome Database (LGD).

Procedure

1. Your primary task in Activity 1 was to determine, on the basis of a rather small amount of data, the structure of the pedigree for your extended family. Follow your teacher's instructions for forming the same work groups in which you participated during Activity 1.
2. Check your pedigree against the pedigree for your family that was generated by the computer and stored in the LGD.
3. Compare the pedigree that appears on the screen to the one that you assembled.
   a. Are there any differences? If there are, discuss the possible reasons for these differences with the other members of your group.

To Access the Pedigree for Your Extended Family

- Click on the *LGD* checkbox.
- Click and pull on *TYPE* to highlight the line that reads *Name*. Release the mouse button.
- Click on *VALUE*. Enter the name of a family member and click on *OK* (or *BEGIN SEARCH*).
- Once the *General Information* screen appears, click and pull on *SCREEN* and highlight *Pedigree*. Release the mouse button and the computer will display the pedigree for the appropriate extended family.

Special Instructions for the Pedigree Screens

The computer-generated pedigree identifies each individual by sample number rather than by name. To see an individual's name, click and hold on the appropriate pedigree symbol. A pop-up box will appear that lists this information, as well as the individual's sample number, sex, age, and genotype.

If you move the cursor off the symbol while holding the mouse button, the box will remain visible on the screen after you release the button. Click on the box to make it disappear.

(**Macintosh Users:** If you wish to go immediately to a specific person's record without going through a search, press the [option] key while you click the mouse on the symbol for the person whose data you wish to see. The database will show you the *General Information* screen for that person. To return to the *Pedigree* screen again, simply click on the *SCREEN* box, highlight *Pedigree*, and release the mouse button.)

NOTE: A diagonal line through a symbol indicates that this person is deceased. It is possible that the pedigree that you drew does not show this detail. Although you may wish to add this information to your pedigree, its absence will not affect your ability to complete the activity.

b. The pedigree on the screen reveals an individual who is not connected to the pedigree and is indicated by an asterisk. This person is the "outlier" referred to in the title of this activity. Discuss the possible reasons for this person being classified as an "outlier."

4. What does the phrase "Incompatible Genotype" mean? Did you take genotypes into consideration when you drew your pedigree? If not, on what basis did you construct your pedigree?

5. Complete Steps I through V on Worksheet 8, *Analyzing the Discrepant Data*, to develop and test some hypotheses about the reason for the apparent discrepancy between the family data and the genotypic data in the LGD. Record your answers in the appropriate places on the worksheet.

6. Follow your teacher's instructions for reporting the results of your investigation to the class.

BLM: Worksheet 8
Analyzing the Discrepant Data

I. Describe the problem.

A. One of the pedigrees in Figure 5 illustrates the portion of the pedigree that is at issue in your family. Examine the pedigree that the computer provides for your family and write the full genotype for each of the individuals involved below his or her name. One genotype has been written in for you already.

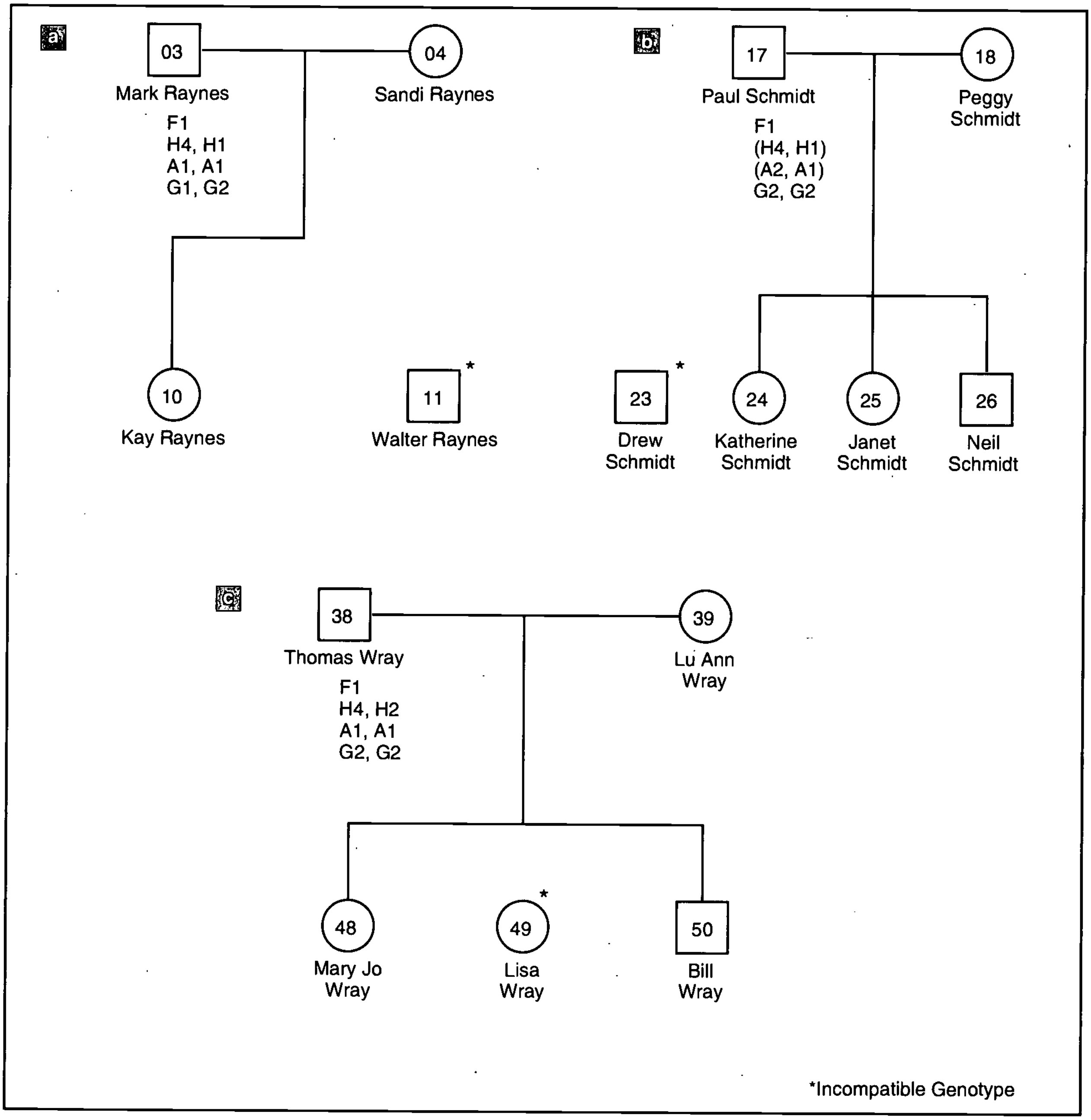


Figure 5 Portions of the pedigree for each of three fictitious families. a. Mota/Raynes/Chen/McCarthy family; b. Jacobs/Schmidt family; c. Thomas/Major/Wray family.

B. 1. Is there an incompatibility in the genotypes that you have listed? Explain.

NOTE: In each case, the genotypes are given so that the gene a person inherited from his or her mother is listed first. For example, a genotype of A1/A2 indicates that the person inherited an A1 allele from his or her mother and an A2 allele from his or her father. When the alleles in question might have been inherited from either parent, the allele symbols are listed in parentheses. Take this into consideration when you determine the nature of the genetic incompatibility that is involved in your fictitious family.

2. If there *is* an incompatibility, does it involve only one gene, or does it involve more than one gene?

3. Does it involve one parent or both parents? Be specific.

II. Collect information that may help you solve the problem.

Your work in Step I should have convinced you that the data in the LGD contain an important discrepancy. Although reported family relationships indicate that this individual is a member of this family, his or her genotype is incompatible.

An important step in solving any problem is to collect relevant information. In this case, for example, it may be important to know something about the genes that are displayed on the pedigree and to understand how the related phenotypes are or are not expressed in the individuals involved.

Take turns following the Steps A and B to collect information about the genes represented in Drew's, Walter's, or Lisa's file.

A. Choose one of the genes to study in more detail. Search the NGD (the model research database) for information about the gene you selected. Enter the information that you retrieve into Table 2.

NOTE: Each of you should choose a different gene to investigate.

Symbol	Gene Name	Definition	Symbols for the Known Allelic Variations and Their Associated Phenotypes

Table 2 Information about the genes represented in Drew's, Walter's, or Lisa's file.

To Locate the File on a Gene in the NGD

- Click on the *NGD* checkbox.
- Click and pull on the *TYPE* box and highlight the line that reads *Gene Symbol*. Release the mouse button.
- Click on the *VALUE* box. The computer will display a list of symbols. Highlight the symbol of the gene for which you wish to search and release the mouse button.
- Click *BEGIN SEARCH* to start the search.
- If necessary, use the scroll bar to read all of the notes.

To Change to the Allelic Variations Screen

- Click and pull on the *SCREEN* box to highlight the line that reads *Allelic Variations*.
- Release the button. The computer will display the new screen.

NOTE: When you first access a gene's file, the computer always will display the *General Information* screen. To move back to this screen, just click and pull on the *SCREEN* box, highlight *General Information*, and release the button.

B. What is Drew's, Walter's, or Lisa's genotype? Based on the information your team has collected about the genes involved, does his or her genotype match his or her phenotype? Explain.

III. Develop hypotheses.

What hypotheses can you propose to explain the discrepant information in the LGD? List your hypotheses below.

IV. Test your hypotheses.

A. Two hypotheses that could explain the discrepancy are sample errors (for example, a lab technician may have confused this person's sample with someone else's) or data error (for example, this person's sequence for this gene might have been entered incorrectly or a lab technician may have made an error as the DNA was sequenced). How could you determine whether there has been a mixup in one of these steps?

1. Whom did you choose to retest and what were the results?

Person(s) Retested	Results

To Retest the DNA Profile for an Individual in the LGD

- Activate the retest function by clicking and pulling on *ACTIVITY* at the upper left of your screen and highlighting the line *Explaining the Outliers, Part B.*
- Your teacher will give you the password to enter. Click on *OK*.
- Select the LGD database.
- Return to the *Pedigree* screen for your family by doing a name search and highlighting the *Pedigree* screen. A button labeled *RETEST* will appear. (The *RETEST* button will appear only on the *Pedigree* screen.)
- Click on *RETEST* and enter the first and last name of the individual whom you wish to retest. Click on *OK* to complete the retest.
- You will see a message reporting the results of the retest. Read the message, then click on *OK* to continue your work.

NOTE: Because of the relatively high cost of DNA testing, the lab will accept only four retest requests from you during this work session. Choose the individuals you wish to retest carefully.

2. Has retesting resolved the discrepancy? Explain.

B. Some of your hypotheses may have stated that Drew, Walter, or Lisa is not the genetic child of one or both of their parents.

1. In theory, what types of evidence available from the LGD would support a hypothesis of adoption?

2. What types of evidence would support nonpaternity?

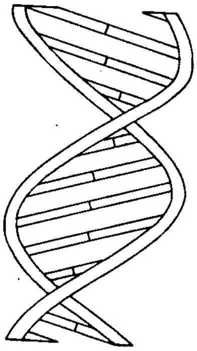
3. What types of evidence would support nonmaternity?

C. How does Drew's, Walter's, or Lisa's genotype fit with those of his or her parents? Explain.

V. State your conclusion.

A. What seems to be the simplest explanation for the observed incompatibility in your fictitious family?

B. How strong is the evidence for your conclusion? Explain.



Activity 3
Genetic Anticipation

Procedure

Imagine that it is now six months after the local health authorities first created the LGD and allowed public searching of a small subset of its data. Joy Major is 21 years old, is working her way through the local community college, and, until yesterday, was thinking about becoming a high school mathematics teacher. She finds school challenging and has to work hard to maintain her grades, but she thinks that this will help her be a good teacher. She has promised herself that *she* will never forget how it feels to be confused.

Two weeks ago, Joy was assigned a genetics project in her biology class. This project involved drawing a pedigree for her immediate and extended family and tracing the inheritance patterns of several genetic traits through this pedigree. Joy had been enjoying this exercise until conversations with her older sister, Leah, revealed that doctors had told Leah that her son's mental retardation likely was inherited. When Joy expressed some surprise at this news, Leah told Joy that when her son was diagnosed, the tests used to trace the inheritance of the condition were hard to interpret. Because of this, Leah had declined further testing for herself and her son and had not told the rest of her family for fear of worrying them unnecessarily.

After Joy talked with Leah, Joy became curious about what scientists might have discovered about the inheritance of fragile X since her nephew was born and she decided to do some research on her own. First, she found the newspaper clipping that your teacher has given you. When she read the article, she became quite upset.

1. Do you see anything in the article that suggests why Joy might be so upset?

HINT: Make sure that you understand what the article says. Answering each of the questions below may help you with this.

a. With what inherited condition is the fragile X gene associated?

b. What is unusual about the manner in which this condition is inherited?

c. What is unusual about the DNA sequence of the fragile X gene, in comparison with many other genes?

d. How many trinucleotide repeats normally occur in the fragile X gene?

e. What is different about the number of these repeats in individuals said to carry a "premutation" for fragile X?

f. How can the number of repeats change as premutations are transmitted from one generation to the next? What are the possible phenotypic consequences of such changes?

2. Follow your teacher's instructions for breaking into small work groups around the computers to discuss Joy's situation more closely. Use Questions 3-7 below to guide your discussion. Consult the databases as required, and be prepared to join the rest of the class in an open discussion by the time your teacher specifies.

3. When Joy first read the clipping, she remembered that fragile X was one of the genes that was recorded in her LGD file. As soon as she got the chance, she looked up her file and then cross-referenced it to the information in the NGD.

   a. What is Joy's genotype for fragile X?

   b. From whom did she inherit the allele that concerns her?

   c. Why may this information have frightened her?

4. After Joy examines her own file, she looks up her siblings' genotypes for the fragile X gene.

   a. What genotype with respect to fragile X does Leah carry?

   b. How does this compare to Joy's genotype? What might this mean for Joy?

   c. Consider Joy's other sister, Anna. What is her situation with respect to fragile X?

   d. What is their brother's genetic status with respect to this gene?

5. Joy also searches the LGD for evidence of the fragile X syndrome in other families in the community.

   a. What does she find?

HINT: You may want to consider searching the LGD by gene symbol to retrieve this information.

   b. How might this information influence Joy's attitude toward these other community members? How might this information affect Joy's feelings about herself?

6. One of Joy's responses to the information in the LGD and the NGD is to wish that she and her family had not agreed to allow their personal files to be made public, *even* for educational purposes. Do you think that she has a point? Why or why not?

7. Examine Joy's genotype with respect to fragile X very closely. What does it tell you? What does it *not* tell you? Based on the data that you have available, what can you say about the probability that Joy carries a fragile X premutation?

8. Follow your teacher's instructions for discussing the following questions as a class.

Questions for Discussion

1. What is the dilemma that faces Joy?

2. What is unusual about the fragile X gene that makes Joy's situation particularly confusing?

3. What role did the rapid increase in scientific understanding play in Joy's dilemma?

4. What role did electronic databases play in Joy's dilemma?

5. Joy attempted to interpret complex genetic information without qualified help. What role did this play in her dilemma? Why is it difficult to define "normal" in this situation?

6. What implications does Leah's son's mental retardation have for her sisters? In what sense are genetic data private and in what sense do we share such data with others? Explain.

7. If Joy consulted a genetic counselor, she would learn, among other things, that there are prenatal (before birth) tests to detect the presence of the fragile X mutation in the developing fetus. How might this information affect Joy as she thinks about her future?

BLM: Worksheet 9

Genetic Anticipation and Trinucleotide Repeats

Scientists Begin to Unravel the Fragile X Mystery

Scientists associated with the Human Genome Project (HGP), a government-sponsored research effort directed at locating all 50,000-80,000 human genes, have reported the discovery of a startling new genetic process that may change the way we think about human inheritance. Called "the expansion of trinucleotide repeats," the process involves a mutation that results in an increased number of repeats of a specific trinucleotide. The mutation also causes the mutated piece of DNA to be even more likely to mutate further. Such mutation further increases the number of repeats of the trinucleotide involved. The discovery of this mechanism helps to explain the phenomenon that geneticists call "genetic anticipation," the observation that certain disorders may become more severe or appear at a younger age in succeeding generations. This pattern of inheritance is unlike the inherited changes that occur in most other types of genes. In other genes, mutations usually are passed on to the next generation with no further change and with little or no change in the severity of the disorder.

One of the genes in which scientists have observed the expansion of trinucleotide repeats is the gene associated with fragile X mental retardation. An abnormality in the fragile X gene is the most frequent cause of *inherited* mental retardation and affects approximately 1/1,500 males and about 1/2,500 females. Until recently, the pattern of inheritance of this condition had baffled scientists. As the name suggests, the gene is located on the X chromosome. Although it is expressed as a dominant trait (that is, only one copy of the altered form of the gene has to be present for mental retardation to occur), both males and females can carry mutated forms of the gene and yet show no signs of mental deficiency. For example, only about half of the female carriers of one fragile X allele show any signs of mental impairment; the rest appear unaffected.

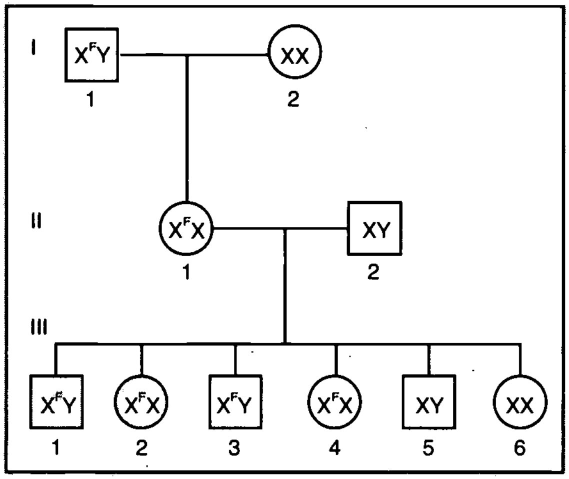


Figure 6 Example of a possible inheritance pattern of the fragile X gene. Individuals with mental retardation are indicated by the gray shading. The normal transmitting male is individual I-1. Notice that his X chromosome carries a mutated form of the fragile X gene (X^F). His daughter (individual II-1) inherits this gene from him, though she usually is not mentally retarded. His grandchildren, who receive the gene from their mother, may or may not be affected depending on whether the gene mutates further during egg production. Because each new conception involves a new egg that may or may not experience further mutation of the gene, all of the outcomes shown (see individuals III, 1-6) are possible.

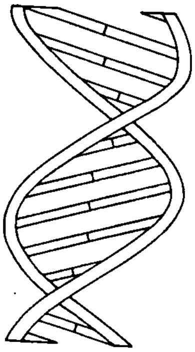
The inheritance pattern of fragile X has been particularly confusing in the case of so-called *normal transmitting males,* who carry mutated forms of the fragile X gene even though they show normal intelligence. Such men may pass these genes on through their daughters (who usually are not mentally retarded) to their grandchildren (who often are). The pedigree in Figure 6 illustrates this inheritance pattern.

(continued)

The breakthrough in understanding the fragile X mystery came after scientists working on the HGP located and partially described the gene responsible. At least one reason for the gene's strange inheritance pattern has to do with a section of bases near one of its ends. In so-called normal genes, this section contains between about 5 and 50 copies of the three-base combination CGG arranged side-by-side without any intervening bases. In normal transmitting males, however, this section shows an increase in the number of repeats. This change in the number of repeats is called a *premutation* because these men may carry between 50 and 200 copies of the repeated unit (CGG), yet show no symptoms.

Daughters who inherit this premutation from their fathers may have normal intelligence, or may show some learning difficulties. All such women, however, are at risk for having mentally retarded children. Apparently, as the gene passes through these women, the size of the repeat area can increase (that is, the gene can mutate even further). Oddly enough, the extent of this increase can be different for each child. This means that a woman who receives a premutation from her father may have some children who carry as many as 1,000 repeats (these children generally are severely mentally retarded) and others who have as few as about 200 (these children may or may not show learning problems).

The discovery of a mutation that makes that gene more likely to mutate further has astonished most investigators. The idea that DNA is not necessarily inherited from our parents in exactly the same form, but can be changed significantly as it passes from parent to child, reminds us that human inheritance involves much more than the transmission of single, unchanging genes from one generation to the next. The determination of the sequence of the fragile X gene also means that scientists have yet another tool that they can use in their efforts to understand fragile X mental retardation more thoroughly.



Activity 4
Who Should Control Information about My Genes?

At the end of Activity 1, your teacher asked you to decide whether to authorize further testing of your fictitious person's DNA. The purpose of this testing was to determine his or her genetic status with respect to the genes for cystic fibrosis (CF), sickle cell disease (SCD), and familial hypertrophic cardiomyopathy (CM).

In this activity, you will consider several options for how to handle the test results.

Procedure

1. Good ethical analysis of science-related issues requires a solid, accurate understanding of the relevant science. Follow your teacher's instructions for forming work groups around the available computers. Consult the NGD for descriptions of the disorders listed above and answer the following questions for each.

   a. What are the medical consequences (if any) for a person who discovers that he or she carries one copy of this allele?

   b. What are the potential reproductive implications (if any) for a person who discovers that he or she carries one copy of this allele?

2. Follow your teacher's instructions for participating in a class discussion about the advantages and disadvantages of handling the test results in several different ways. First, you will explore the following three options:

   □ report only the anonymous, combined data (Option 1)
   □ report each set of individual results only to the person tested (Option 2)
   □ announce the individual results openly or publicly to all (Option 3)

Your class will vote after the discussions about Options 1 and 3.

3. Return to your small groups to consider a final option for disseminating the individual test results (Option 4) by answering the following question:

   Should the individual test results from GeneTest, Inc. be entered into the LGD?

   Use what you know about the genes in question, the arguments you and your classmates already have considered, and your experience with computerized databases as a basis for your discussion and decision.

4. Follow your teacher's instructions for sharing your list of "yes" or "no" reasons with the rest of the class.

5. Consider both sets of reasons for a moment. What do you believe to be the most compelling

reason or reasons on each side of the issue? Explain your answer.

6. Are these two sets of reasons equally compelling? Why or why not? Explain your answer.
7. Follow your teacher's instructions for deciding whether the individual test results should be entered into the LGD.
8. Can you think of any other situations in which access to personal genetic information in a registry might create ethical dilemmas?

Homework Assignment

In Activity 4, you considered a series of questions about how to handle the new genetic information that was gathered about your fictitious person as a result of additional DNA testing. Your discussions culminated in a decision about whether to enter those data into a registry database.

Your vote on this question likely was influenced by your experience with the databases in Activities 1-3. Those of you who voted to enter the data probably considered the advantages of having such information stored and accessible for research and health-care purposes. Those of you who voiced objections probably considered the disadvantages of having personal and potentially sensitive information accessible to those who should not have that information.

Many registry databases already exist in the United States. Some of these databases store actual genetic information; others store medical information from which one can infer various types of genetic information. Many DNA databanks also exist. These are databases that store actual DNA samples or that store tissue samples from which DNA can be extracted. The individuals responsible for electronic databases and databanks usually adhere to relatively strict criteria about access to and use of these data. Many private citizens and organizations, however, believe that we need clearer and more uniform guidelines to protect the privacy of genetic information.

In Activity 5, your class will act as a committee of elected legislators to discuss and to recommend such guidelines. In preparation for your work as a legislator, read the background material on Worksheet 10, *Genetic Registries, Information, and Privacy*, and answer the study questions provided.

BLM: Worksheet 10
Genetic Registries, Information, and Privacy

Background Information on Current Privacy Laws in the United States

At present in the United States, there are no special federal laws or regulations in place that protect the privacy of genetic information. Instead, genetic information is treated like medical information and is subject to the same restrictions that apply to one's health records.

Genetic Information Is Different from Medical Information: There are many individuals in the United States who believe that it is time for society to think critically about the rules that govern the collection, storage, and distribution of genetic information. These individuals argue, in part, that the increase in our understanding of and ability to study the human genome has made the issues of collecting, storing, and using genetic information more urgent and more important than ever before.

These individuals also argue that genetic information has a number of characteristics that make it distinct from medical information and that society needs to recognize these differences and create appropriate mechanisms to protect it. For example, unlike some health information, one's genetic make-up is completely beyond one's control. Genetic information also is thought by many to have more power to predict one's future health than most other types of health assessments. Genetic information about one person has implications for other people, especially those in one's immediate and future families, and information about one's genetic risk can be psychologically and socially stigmatizing (for example, it can lead to people feeling or being treated as though they are not "normal"). Finally, people who argue for tighter controls on the collection and use of genetic information also point out that DNA is a very stable (long-lasting) chemical, and that samples taken for one purpose now may be used for other, perhaps unauthorized, purposes later.

State Regulations that Protect Genetic Information: Although there are no federal restrictions in place with respect to genetic information, several states have created laws and regulations that place some restrictions on access to and use of such information. For example, by mid-1995, twelve states (Alabama, Arizona, California, Colorado, Florida, Louisiana, Maryland, Montana, North Carolina, Ohio, Tennessee, and Wisconsin) had prohibited the use of genetic information by insurers.

Most of these restrictions apply only to one or two specific genetic conditions or to people who carry certain genetic disorders but who remain unaffected by them. In 1992, Wisconsin passed a law that prohibits health insurers from requiring or requesting individuals to take a DNA test, to reveal whether they ever have undergone a DNA test, or to disclose the results of such tests. The law also prohibits insurers from using DNA test results to determine rates or other aspects of health-insurance coverage. Ohio's law (passed in 1993) calls for a ten-year moratorium on the use of information from genetic tests, and Colorado's law (passed in 1994) prohibits the health, group, and long-term care insurance industries from seeking, keeping, or using genetic testing information to make decisions about insurance coverage or for any other reasons that are not directly related to a patient's care or treatment.

Arguments For and Against Federal Action: Despite the work that already has been done to establish state regulations to govern the collection and use of genetic information, there are many individuals who believe that it is time for the federal government to act. Such people argue that it is difficult to achieve comprehensive or uniform control when the laws are written and enforced at the state level. On the other hand, others argue that federal regulations cannot address the real heart of the question because many genetic registries and databanks are in private hands and are not governable by federal legislation.

Constitutional and Common-Law Protections to Privacy: Underlying the discussions about access to and use of genetic information are some basic understandings about our right to privacy as protected by the Constitution and certain common-law commitments. For example, according to George Annas (1993)[1], in the United States, there are four major senses of privacy (or ways in which our privacy is protected). The central one, protected by the Fourteenth Amendment, limits the ability of the government to interfere with intimate, individual decisions (such as those involving reproduction). The second and third sense of privacy, also protected by the Constitution, protect certain relationships (such as husband and wife, health-care provider and patient) and certain places (such as the home) from government intrusion. The fourth sense of privacy, the common-law sense of privacy, protects our right "to be left alone," that is, our right to refuse to share personal information with others.

How do these senses of privacy apply to genetic information? First, genetic diseases differ from contagious diseases in that the spread of genetic disease is controlled by reproductive decisions and actions, as opposed to immunizations, antibiotics, sanitation, and quarantine. Thus, genetic information might be protected under our constitutional right to be protected from interference in intimate, individual decisions. Second, genetic information is family information that often is discovered in the health-care provider and patient relationship and might be protected under the constitutional protections that cover that special relationship. Finally, genetic information is personal and unique in that it affects our self-identity and might be protected under our common-law right to refuse to share personal information with others.

Background Case

The following is a true report submitted by a clinical geneticist in the United States caring for individuals diagnosed with phenylketonuria (PKU). We have changed the child's name to protect her privacy. The case study is presented as an indication of the need to consider public policy concerning genetic information.

"[Jennifer] is an 8-year-old girl who was diagnosed as having PKU at 14 days of age through the newborn screening program in Connecticut. A low phenylalanine diet was instituted at that time.

Jennifer's growth and development have been completely normal. Routine developmental assessments done at 26 weeks, 53 weeks, and 54 months revealed that Jennifer had solid skills that were appropriate for her age, and in many instances, skills that were above what would be expected for her age. The child continues to be developmentally normal and healthy.

Recently, Jennifer and her family lost their medical insurance (through Eucare) that had provided for her treatment. Her father changed jobs and the new insurance company (Guard) notified the family that Jennifer was considered to be a high-risk patient because of her diagnosis and, therefore, was ineligible for insurance coverage under their group plan.

When Jennifer's parents investigated to determine why Jennifer was denied health insurance, they discovered that the problem began when the family had been denied supplemental insurance coverage from WholeCare. As a matter of standard practice, most insurance companies send information concerning clients who are deemed uninsurable to Central Information Systems, Inc. (CIS), a company that stores medical, life, and disability insurance information in a registry that is open to insurance companies for a nominal fee. Following the job change, Guard (the family's present insurer) ran an insurance check on the new employee and his family through CIS and discovered Jennifer's earlier diagnosis. It was through the information stored in this database that Guard decided to deny coverage to 8-year-old Jennifer.

Jennifer currently is being covered at the expense of the family, but this is a temporary solution at best. The family has written to the agency that administers the group insurance plan to obtain further details about the decision to deny coverage and also plans to write to the chairman of the large corporation for which Jennifer's father works. All of the information also will be submitted to the state insurance commissioner." [2]

Study Questions

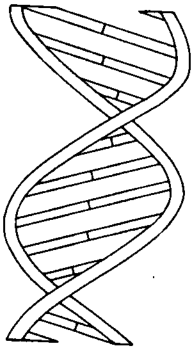
Reflect on what you learned from the activities that you already have completed and from the background information on Worksheet 10. Think as well about the PKU case.

a. What are some *advantages* of laws that place strict protections on the privacy of genetic information?

b. What are some *disadvantages* of such laws?

NOTES:

1. Annas, G. (1993, November). Privacy rules for DNA databanks. *Journal of the American Medical Association* 270 (19):2346-2350.
2. Revised with permission from: Billings, P.R., Kohn, M.A., deCuevas, M., Beckwith, J., Alper, J.S., and Notowicz, M.R. (1992). Discrimination as a consequence of genetic testing, *American Journal of Human Genetics* 50:478-479.



Activity 5
Making Public Policy

Imagine that you are an elected federal legislator and a member of the legislature's Committee on Genetic Information in Registry Databases. Your first task is to elect a student to serve as the committee chairperson; this individual will guide today's discussion and conduct the concluding committee vote.

Three of your legislative colleagues who are not on the committee have submitted separate recommendations for your consideration. These recommendations are listed below.

Recommendation A:

It is premature to act. There should be no *new* laws developed at this time about the release or the use of genetic information from registry databases.

Recommendation B:

There should be laws that limit an individual's right to privacy with respect to the data in genetic registries when release of the data can be shown to be in the best interests either of that individual or of the community as a whole.

Recommendation C:

There should be laws that guarantee an individual's right to privacy with respect to the data in genetic registries.

To ensure that each of these recommendations is considered carefully, the committee will form three subcommittees. Each subcommittee will discuss one of the recommendations, focusing particularly on how policies consistent with that recommendation would work in the case that you considered in preparation for these discussions. After analyzing its assigned recommendation, each subcommittee will report its findings to the full committee. The committee then will decide by majority vote which recommendation it will present to the legislature for further action.

PROCEDURE

1. Follow your teacher's instructions for meeting in your subcommittee.
2. Your teacher will provide you with a worksheet that states the recommendation that your subcommittee will discuss and that outlines the questions that you will be asked to answer when you report to the full committee. Use the background information on current privacy laws in the United States, the case, and your answers to the study questions that you completed for homework as resources to help you analyze your assigned recommendation. You will have about 15 minutes to complete your deliberations.

3. Choose a spokesperson to report the results of your subcommittee's analysis to the full committee. Follow the committee chairperson's instructions for completing and discussing the reports from the subcommittees.

4. Remember that you, as well as your colleagues, eventually will be asked to vote for the recommendation that you believe is most likely to lead to actions that will best address questions about the protection of data in genetic registries. In preparation for voting, follow your chairperson's instructions for participating in a general discussion of all of the recommendations. The questions below will help you organize your thinking about this issue.

   □ Which recommendation, if any, do you think works well in the PKU case? Explain your reasoning.

   □ Which recommendation, if any, do you think would work well in most cases? Explain your reasoning.

   □ What questions do you have about these recommendations that have not been raised or addressed? What alternative recommendations do you have to offer?

5. Follow your chairperson's instructions for casting your vote.

   NOTE: You are *not* required to vote for the recommendation that you analyzed. Instead, you should vote for the recommendation that you think would lead to the best outcome.

6. Reasonable people can come to very different decisions about a controversial question or issue. What does this suggest about the process of developing public policy?

BLM: WORKSHEET 11
Analyzing Recommendation A

Recommendation A:

It is premature to act. There should be no new laws developed at this time about the release or the use of genetic information from registry databases.

Notice that the goal of Recommendation A is to preserve all options and to allow more time for discussion. The questions below will help you think about how this recommendation would work in the case that you read in preparation for this discussion.

Consider the PKU case. Would this recommendation have allowed the insurance company to access and use the data in this manner? (circle one) yes no

Give some reasons that this would be a bad outcome. Which reason do you think is the most important? (check one)

Check Reasons

_____ ______________________________________

_____ ______________________________________

_____ ______________________________________

Give some reasons that this would be a good outcome. Which reason do you think is the most important? (check one)

Check Reasons

_____ ______________________________________

_____ ______________________________________

_____ ______________________________________

Discuss the two reasons that you have identified as being the most important. Place a second check in front of the reason that your subcommittee finds the most compelling.

Do you think that Recommendation A works well in the PKU case? (circle one) yes no

Do you think that Recommendation A would work well in most cases? Explain your answer and support it with other specific examples of situations in which you think that it would or would not work well.

BLM: WORKSHEET 12
Analyzing Recommendation B

Recommendation B:

There should be laws that limit an individual's right to privacy with respect to the data in genetic registries when release of the data can be shown to be in the best interests either of that individual or of the community as a whole.

Notice that the goal of Recommendation B is to protect individual and societal health and well-being. The questions below will help you think about how this recommendation would work in the case that you read in preparation for this discussion.

Consider the PKU case. Would this recommendation have allowed the insurance company to access and use the data in this manner? (circle one) yes no

Give some reasons that this would be a bad outcome. Which reason do you think is the most important? (check one)

Check Reasons

_____ ____

_____ ____

_____ ____

Give some reasons that this would be a good outcome. Which reason do you think is the most important? (check one)

Check Reasons

_____ ____

_____ ____

_____ ____

Discuss the two reasons that you have identified as being the most important. Place a second check in front of the reason that your subcommittee finds the most compelling.

Do you think that Recommendation B works well in the PKU case? (circle one) yes no

Do you think that Recommendation B would work well in most cases? Explain your answer and support it with other specific examples of situations in which you think that it would or would not work well.

BLM: Worksheet 13
Analyzing Recommendation C

Recommendation C:

There should be laws that guarantee an individual's right to privacy with respect to the data in genetic registries.

Notice that the goal of Recommendation C is to protect privacy. The questions below will help you think about how this recommendation would work in the case that you read in preparation for this discussion.

Consider the PKU case. Would this recommendation have allowed the insurance company to access and use the data in this manner? (circle one) yes no

Give some reasons that this would be a bad outcome. Which reason do you think is the most important? (check one)

Check Reasons

_____ ____

_____ ____

_____ ____

Give some reasons that this would be a good outcome. Which reason do you think is the most important? (check one)

Check Reasons

_____ ____

_____ ____

_____ ____

Discuss the two reasons that you have identified as being the most important. Place a second check in front of the reason that your subcommittee finds the most compelling.

Do you think that Recommendation C works well in the PKU case? (circle one) yes no

Do you think that Recommendation C would work well in most cases? Explain your answer and support it with other specific examples of situations in which you think that it would or would not work well.

BLM: WORKSHEET 14
Instructions for the Committee Chairperson

As chairperson of the committee, your tasks are to

- □ call the subcommittees together after they have completed their deliberations
- □ organize the process of receiving oral reports from the subcommittees
- □ invite discussion about the recommendations
- □ call for the final vote

The information provided below will help you organize your work. The numbers are keyed to the numbers used in the student procedures.

3. **Call the subcommittees together after they have completed their deliberations.**

   After the subcommittees have spent about 15 minutes discussing their assigned recommendations, call the full committee to order. When your classmates have reassembled, briefly explain how you will conduct the subcommittee reporting and the voting.

   Use the overhead transparency that your teacher will give you to record the main points of each subcommittee's report. Begin by asking the spokesperson for the subcommittee that analyzed Recommendation A to answer the questions listed on the transparency and continue until you have heard reports on all three of the recommendations.

   If more than one subcommittee analyzed a particular recommendation, ask each of the spokespersons involved to comment on his or her group's analysis of each item listed on the transparency. If the groups give similar answers, enter that answer on the transparency and continue with the next item. If the groups disagree or give dissimilar responses, ask the spokesperson for each group to explain why his or her group answered as it did. If the groups reach consensus (for example, if one group changes its answer), enter that response and go on to the next item. If the groups do not reach consensus, indicate the disagreement on the transparency and continue to the next item.

   Invite discussion of each of the subcommittee reports by pausing after the report is complete and asking the class whether anyone has questions about the recommendation or whether anyone would like to raise other issues or arguments that the subcommittee did not appear to consider. Do not attempt to answer these questions or respond to these issues or arguments yourself. Instead, refer them back to the students who served on the subcommittee(s) that analyzed that recommendation and invite them to comment. Invite other students in the class to comment as well.

4. **Invite more general discussion about all of the recommendations.**

   After the subcommittees' reports are complete, invite the class to comment more generally on the recommendations. Remind the class that each student will be asked to vote for the recommendation that he or she believes is most likely to lead to actions that will best address questions about the protection of data in genetic registries.

   You may wish to organize this discussion by asking the students to respond directly to each of the questions listed in the student text.

   If the students have difficulty imagining cases in which one or more of the recommendations would not work well, you may choose to read the following situations to the class, pausing after you read each one to ask how well they think each recommendation would work in that case and why.

   A. The Arapahoe Women's Health Clinic has a large registry containing genetic data on its patients. Included in these data is information about genetic susceptibility to hypertension during pregnancy, as well as other personal information. Hypertension during pregnancy causes serious health risks for the woman and can lead to a life-threatening stroke. STB

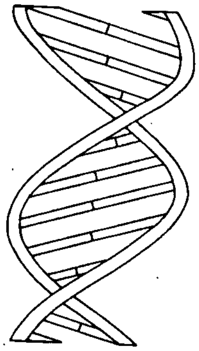
Pharmaceuticals is marketing a new drug that researchers have shown is particularly effective and safe in preventing hypertension during pregnancy. STB Pharmaceuticals has requested access to the registry to target its marketing efforts more specifically to women at risk for this condition.

B. Females who are carriers of fragile X often have learning disabilities. Amalgamated School District #2 wants access to registry data to identify and track these students so that they can receive any academic help they may require.

C. Tricia Payne, a high school student, is looking for a simple research question to investigate for her science fair project. She has requested access to registry data to identify those students who carry particular allelic variations of the alpha hemoglobin gene. She would like to test whether there is any relationship between an individual's genotype for alpha hemoglobin and his or her musical ability.

5. **Call for the final vote.**

Call the final vote by asking for a show of hands of students in favor of presenting Recommendation A, Recommendation B, or Recommendation C to the legislature. Each student should vote for only one recommendation; this does *not necessarily* have to be the recommendation that he or she helped to analyze. Tally the results on the board.



Extension Activity HGP Data and Evolutionary Biology

Part A
DNA, Political Assassination, and World History

Genomic databases can include DNA sequence data from living individuals or from persons long dead. The sequence data for this activity, which you will retrieve from the research database, include actual data from DNA taken from nine human skeletons found in a shallow grave in 1991.

The DNA used for this analysis is mitochondrial DNA, that is, DNA taken from the mitochondria of cells. Recall from your study of cells and of energy production in living systems that mitochondria are contained in the cytoplasm of the cell. Their primary function is the production of ATP, the chief source of energy in living systems. Mitochondria have their own DNA, apart from the DNA found in the nucleus. The complete sequence of human mitochondrial DNA, or mtDNA, as it is called, was determined in 1991. The sequence of 16,569 bases is included as part of international DNA sequence databases.

All human cells—including ova and sperm—contain mitochondria. Biologists have observed, however, that mitochondria are transmitted only through females—to their sons and daughters. Propose a hypothesis to explain this observation.

HINT: Consider the location of mitochondria in sperm cells. What happens to a sperm cell during fertilization?

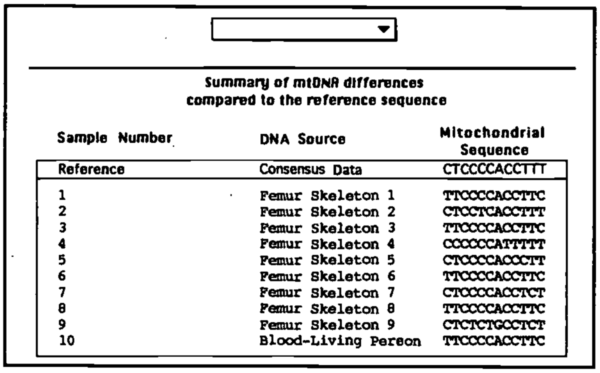
In addition to being maternally inherited, some regions of the mitochondrial genome mutate readily, so there is enough variation to distinguish families from one another.

Procedure

1. The first screen for this activity is shown in Figure 7. Note that it displays the following:

   □ reference data for a 12-base sequence of human mtDNA,
   □ mtDNA sequences from nine skeletons recovered from a shallow grave in 1991, and
   □ an mtDNA sequence from a living person.

2. To make it easier to find sequences that are identical, it is convenient to use a dot (.) to replace any base that is identical to the base that occurs at the same position in the reference sequence. Bases that do not match remain listed as letter symbols. Follow the instructions below to insert dots in any positions in which the bases are identical to those in the reference sequence. What do you notice about the sequences?



Summary of mtDNA differences compared to the reference sequence

Sample Number	DNA Source	Mitochondrial Sequence
Reference	Consensus Data	CTCCCCACCTTT
1	Femur Skeleton 1	TTCCCCACCTTC
2	Femur Skeleton 2	CTCCTCACCTTT
3	Femur Skeleton 3	TTCCCCACCTTC
4	Femur Skeleton 4	CCCCCCATTTTT
5	Femur Skeleton 5	CTCCCCACCCTT
6	Femur Skeleton 6	TTCCCCACCTTC
7	Femur Skeleton 7	CTCCCCACCTCT
8	Femur Skeleton 8	TTCCCCACCTTC
9	Femur Skeleton 9	CTCTCTGCCTCT
10	Blood-Living Person	TTCCCCACCTTC

Figure 7 Sample screen from *HGP Data and Evolutionary Biology, Part A.*

To Replace Identical Bases with Dots

Click and pull on the downward-pointing triangle that appears in the box at the top of the screen. Highlight the line that reads *Replace with Dots* and release the mouse button. The computer will replace any base that is identical to the reference data with a dot.

To Move the Sequences

For Macintosh Computers

- Click and drag the sequence you wish to move. Release the mouse when the cursor reaches the desired position.

NOTE: The sequence will not appear to move at first. Instead, an arrow will appear to the left of the sequence you are dragging. This arrow indicates that the sequence *will* move when you release the mouse button.

For Windows Computers

- Click on the sequence you wish to move. This will highlight it.
- Click on the Up or Down button to move the highlighted sequence to the desired position.

3. Work with your team members to cluster the sequences that are most similar.

NOTE: You cannot move the reference sequence.

4. Among the recovered skeletons are those of an adult female and three children. Based on this information, the sequence data, and the method of transmission of mtDNA, propose some hypotheses about the skeletal remains.

5. As a class, read and discuss the following information:

The official records of the fate of the Russian royal family (the Romanovs) are scant, but it is known that prior to their deaths they were held prisoner in Ipatiev House at Ekaterinburg, in the Ural Mountains of Central Russia (Figure 8). It is believed that shortly after the night of 16 July 1918, Tsar Nicholas II; his wife, Tsarina Alexandra; their four daughters, Olga, Tatyana, Maria, and Anastasia; and their only son, Alexei, were herded into the cellar together with three of their servants and the family doctor, Eugeny Botkin. They were shot by the Bolshevik firing squad, although a number of the victims were allegedly stabbed to death when gunfire failed to kill them. The bodies were stripped and placed onto a truck with the intention of disposing of them down a mine shaft. However, the truck developed a mechanical fault during the journey; the Bolsheviks placed the victims into a hastily dug pit in a road, and to hinder identification, sulfuric acid was thrown into the open grave. After the bodies were covered, a truck was driven backwards and forwards over the site to flatten the area.

This account is supported by forensic evidence collected in 1918–19 by Nikolai Sokolov (a White Russian monarchist investigator) whose seven-volume dossier has become the basis for the historical version of the fate of the Romanovs. However, the version of events described above has never been positively verified.

After referring to archival materials and photographs, which gave an indication of a burial site, two Russian amateur historian investigators, Gely Ryabov and Alexander Avdonin, announced that they had discovered a communal grave approximately 20 miles from Ekaterinburg. Consequently, the Russian government authorized an official investigation coordinated by the chief forensic medical examiner of the Russian Federation. The grave consisted of a shallow pit (less than 1 m deep) and contained human skeletal remains. Many of the bones were badly damaged, but it was possible to identify nine corpses. All of the

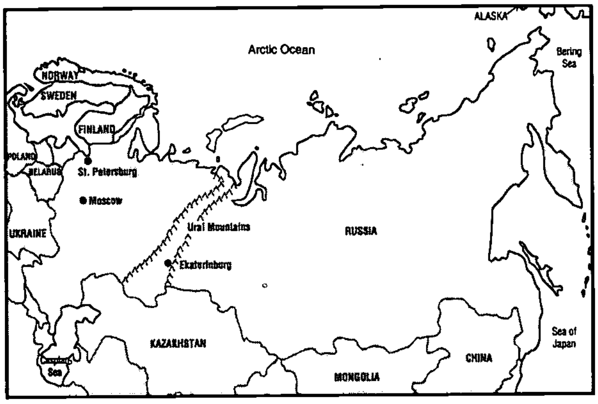


Figure 8 Map of the former Soviet Union.

skeletons showed evidence of violence and mistreatment before death. Some of the skulls had bullet wounds; bayonet marks were also found. Facial areas of the skulls were destroyed, rendering classical facial identification techniques difficult.

6. Follow the instructions below to evaluate the clusters and discover the origin of the samples. Do these data support your hypotheses? Do the data prove your hypotheses?

> **To Evaluate the Clusters**
>
> Click and pull on the downward-pointing triangle that appears in the box at the top of the screen. Highlight the line that reads *Evaluate Clusters* and release the mouse button.

Homework

Complete the exercise described on the worksheet your teacher provides.

Part B
DNA Sequences and Life's History

Procedure

Biologists use other types of sequence data to study evolutionary relationships. The next part of this activity, for example, asks you to work with amino acid sequences.

Amino acid sequences and DNA sequences are like molecular time machines that allow biologists to get a glimpse of life's history on earth. Amino acid sequences, however, allow biologists to "look back" further in time than do DNA sequences. Propose an explanation for this fact.

HINT: How many codons are there? How many amino acids are there?

1. Work with your team members to retrieve a set of amino acid data from the computer. Follow the instructions below to bring up the screen that reads *Amino Acid Sequences for Cytochrome Oxidase.*

> **To Retrieve the Data Required for Part B**
>
> - Click on *ACTIVITY* at the upper left of your screen and highlight the line *HGP Data and Evolutionary Biology, Part B.* Release the mouse button.
> - Enter the password your teacher gives you and click on *OK.*

2. This screen shows the amino acid sequences for a portion of the protein cytochrome oxidase for six different species.

   Note that each amino acid has a one-letter abbreviation. It is not critical that you know which letter represents which amino acid, but that information is included in Table 3.

3. Note that the data are unidentified at this point. Again, work with your team to replace identical amino acids with dots and to move the sequences so that the most identical sequences are grouped together.

4. Follow the steps below to construct a hypothetical evolutionary tree.

   a. First, focus on major groups. Do the sequences cluster into a few major groups of

Table 3 Key to abbreviations for amino acids.

A = alanine	I = isoleucine	R = arginine
C = cysteine	K = lysine	S = serine
D = aspartic acid	L = leucine	T = threonine
E = glutamic acid	M = methionine	V = valine
F = phenylalanine	N = asparagine	W = tryptophane
G = glycine	P = proline	Y = tyrosine
H = histidine	Q = glutamine	

similar sequences? How many groups do you see? Which samples belong to each group?

b. Note that the computer has not identified the species from which the sequence data were derived. Follow the instructions below to retrieve the list of species.

To Retrieve the List of Species

- Click and pull on the downward-pointing triangle that appears in the box at the top of the screen.
- Highlight the line that reads *List Species* and release the mouse button.

c. Based on your knowledge of the species and the data from the computer, work as a team to construct a proposed evolutionary tree. Use Worksheet 18, *Constructing an Evolutionary Tree from Amino Acid Sequence Data*, to help you.

5. When your team has completed the worksheet, evaluate the clusters by clicking and pulling on the downward-pointing triangle that appears in the box at the top of the screen. Highlight the line that reads *Evaluate Clusters* and release the mouse button. The computer will place the species' common names next to their associated sequences. Revise your tree, if you think this new information justifies such a revision.

BLM: WORKSHEET 15
Matching DNA Sequences from Frog, Chicken, Goat, Cow, and Chimpanzee

The DNA sequences shown at the bottom of the page are portions of the beta hemoglobin gene from five different species. Your task is to cut out the sequences and line them up (one under another) so the sequences match to the greatest degree possible.

1. Cut out the sequences and align them in the order you think most appropriate.

2. Tape the sequences to a second sheet of paper.

3. Complete the chart below:

	Frog	Chicken	Goat	Cow	Chimp	Human*
Frog						
Chicken						
Goat						
Cow						
Chimp						
Human*						

Table 2 Numbers of differences in DNA sequences. *Your teacher will give you this sequence when you discuss the data in this table.

4. Take the taped sequences and this worksheet to your next class for additional work.

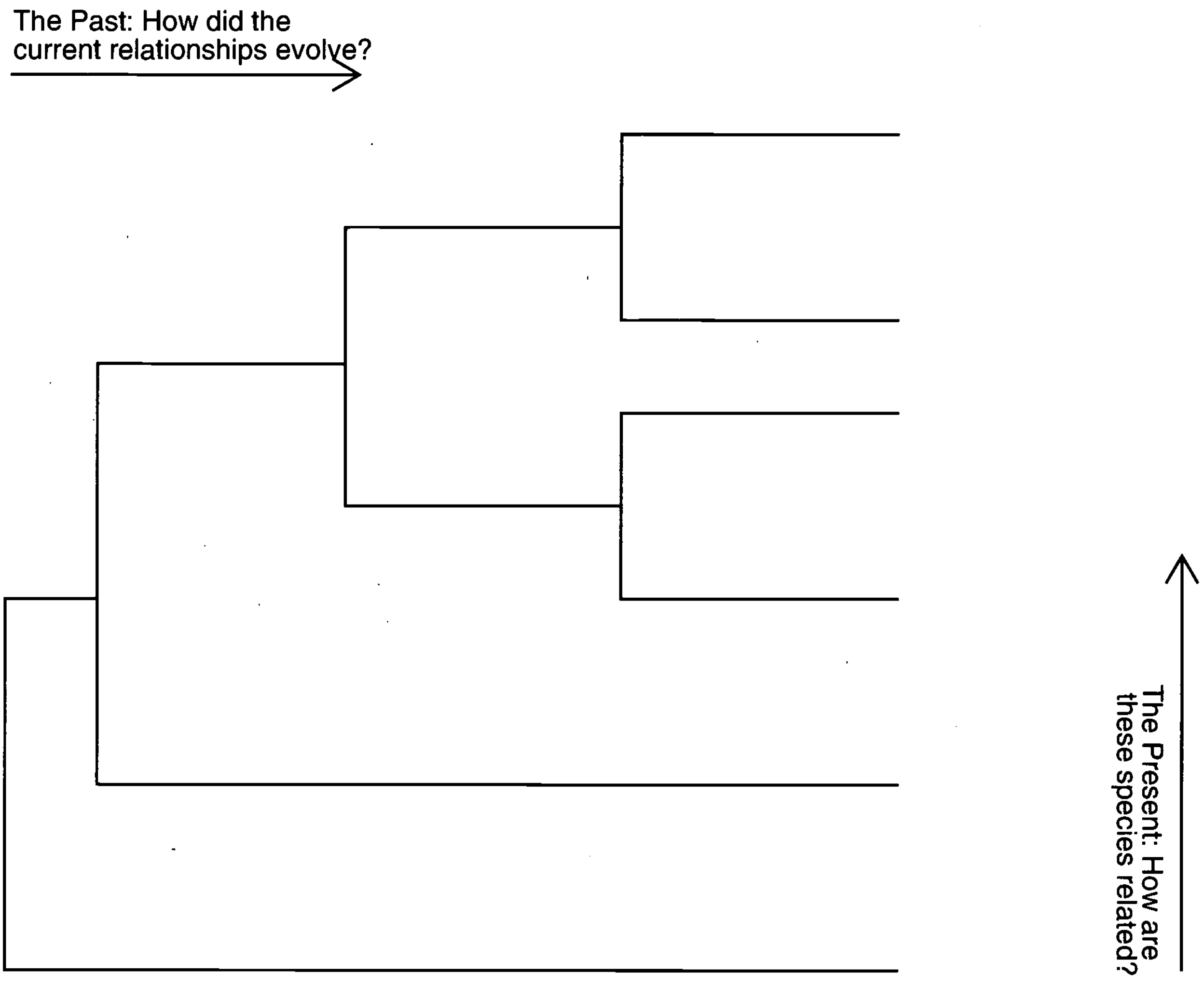
Species	Hemoglobin Sequences
cow	GCTGCACTGT GATAAGCTGC ACGTGGATCC TGAGAACTTC
chimpanzee	GCTGCACTGT GACAAGCTGC ACGTGGATCC TGAGAACTTC
chicken	ACTGCATTGT GACAAGCTGC ATGTGGACCC CGAGAACTTC
goat	GCTGCACTGT GATAAGCTGC ACGTGGATCC TGAGAACTTC
frog	GAAGCACGCT GAGGAACTCC ACGTGGACCC TGAAAACTTC

BLM: WORKSHEET 16
DNA Sequence Data for the Beta Hemoglobin Gene in Humans

Human GCTGCACTGT GACAAGCTGC ACGTGGATCC TGAGAACTTC

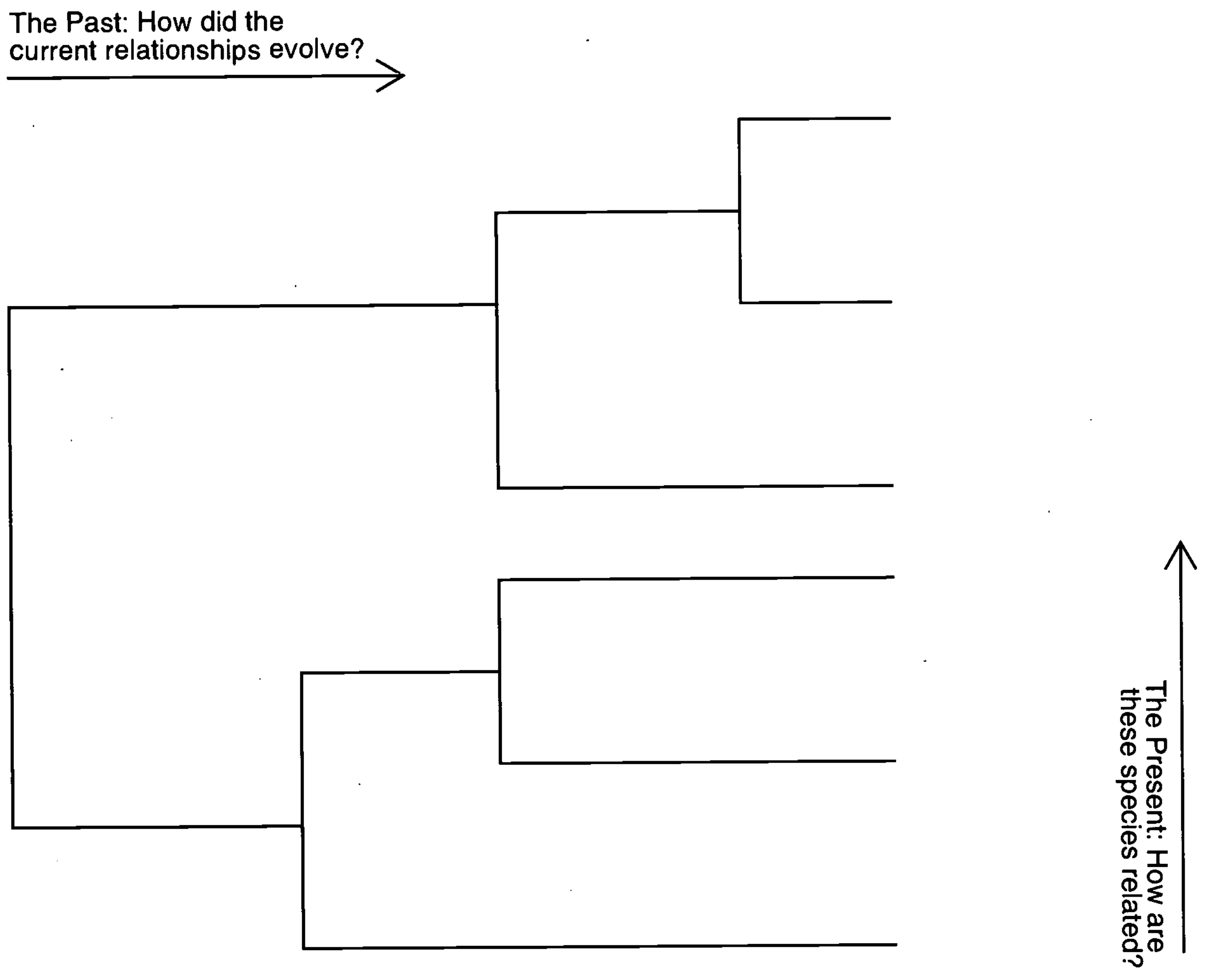
BLM: Worksheet 17
Constructing an Evolutionary Tree from DNA Sequence Data

Use the data that you collected about the number of differences between the frog, chicken, goat, cow, chimpanzee, and human sequences to complete the evolutionary tree below.


The Past: How did the current relationships evolve?
The Present: How are these species related?

BLM: Worksheet 18
Constructing an Evolutionary Tree from Amino Acid Sequence Data

Use the three major groupings (clusters) that you discovered in the amino acid data in the computer and the species' names to place the six species on the tree below.


The Past: How did the current relationships evolve?
The Present: How are these species related?

Evaluation Form for
The Human Genome Project: Biology, Computers, and Privacy

Your feedback is important to us. Please take a few minutes to complete and return this form after you have used the module.

BSCS has developed and distributed this module free of charge with the assistance of the Office of Environmental Health Effects, United States Department of Energy and the University of Iowa Genome Center. Your answers to the following questions will help us assess the effectiveness of the module and will help guide the development and distribution of subsequent, related programs.

1. Please comment on the distribution of the module.

   a. When did you receive this module? Approximate date ______________________________

   b. Was the address on the mailing label correct? yes _____ no _____
   If not, please provide the correct information on the last page of this form.

   c. Do you prefer to receive such material at home ______, at school ______, or other ______________
   ____?

2. Please rate the background materials for the teacher.

	not helpful				very helpful
a. Module-at-a-Glance	1	2	3	4	5

   comments:

b. Section I: What Is the Human Genome Project?	1	2	3	4	5

   comments:

c. Section II: The Science and Informatics of the Human Genome Project	1	2	3	4	5

   comments:

	not helpful				very helpful
d. Section III: Ethical and Public-Policy Dimensions of Research Databases and Registries	1	2	3	4	5
comments:					
e. Implementation Support	1	2	3	4	5
comments:					
f. Glossary	1	2	3	4	5
comments:					
g. References	1	2	3	4	5
comments:					

3. What are the major strengths of this module?

4. What are the major weaknesses of this module?

5. Which version of the software did you use? Macintosh ______ Windows ______. Please provide specific feedback about the software.

6. Please comment on the student activities.

   At what grade level(s) did you use the activities? ______________________________

   With how many students? ____________________

Please rate each aspect of each activity on a scale of 1 (strongly disagree) to 5 (strongly agree).

	Students understood this activity.	Students enjoyed this activity.	The ideas in this activity are important for students to learn.
Introductory Activity: The HGP and Electronic Databases	_____	_____	_____
Activity 1: Genetic Registries	_____	_____	_____
Activity 2: Explaining the Outliers	_____	_____	_____
Activity 3: Genetic Anticipation	_____	_____	_____
Activity 4: Who Should Control Information about My Genes?	_____	_____	_____
Activity 5: Making Public Policy	_____	_____	_____
Extension Activity: HGP Data and Evolutionary Biology	_____	_____	_____

7. Will you use this module again? yes ______ no ______
   Please comment on your decision.

8. Did you use the first BSCS genome module, *Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy?* yes ______ no ______
   If yes, please provide feedback that you think will help us if we revise that module.

9. In early 1997, BSCS will distribute a third genome-related module free of charge, *Changing Concepts of Inheritance: Genetics and the Methods of Science*. Please indicate whether you wish to receive this module.
   yes ______ no ______

10. BSCS hopes to develop and distribute free of charge a fourth genome-related module, tentatively titled *Genes, Environment, and Human Behavior*. Would you likely use such a module in your classes?
    yes _____ no ______

Thank you for your feedback. Please return this form to BSCS, Attn: Dee Miller, Pikes Peak Research Park, 5415 Mark Dabling Blvd., Colorado Springs, Colorado 80918-3842.

Name ______________________________

School ______________________________

Mailing Address ______________________________

____

Phone ______________________________

FAX ______________________________

e-mail Address ______________________________

☆ U.S. GOVERNMENT PRINTING OFFICE:1996-514-675

BSCS

The Human Genome Project: Biology, Computers, and Privacy

A Free Instructional Module for the Biology Classroom

Includes	□	timely background information for the teacher
	□	seven classroom activities
	□	implementation support
	□	software (Windows and Macintosh)
Developed by	□	BSCS
Supported by	□	The U.S. Department of Energy
Distributed with	□	The U.S. Department of Energy
the assistance of	□	University of Iowa Genome Center

SE061480

ERIC®

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Human Genome Project: Biology, Computers, and Privacy

Author(s): BSCS

Corporate Source: Laura Engleman

Publication Date: 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Laura Engleman	Printed Name/Position/Title: LAURA ENGLEMAN PR MANAGER	
Organization/Address: 5415 Mark Dabling Blvd Colorado Springs, CO 80918	Telephone: (719) 531-5550	FAX: (719) 531-9104
	E-Mail Address: lengleman@bscs.org	Date: 7/20/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC/CSMEE
1929 Kenny Road
Columbus, OH 43210-1080

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.