AGENDA

Council Scientific Planning Subcommittee
January 13-14, 1998
Bethesda Marriott Hotel
5151 Pooks Hill Road
Bethesda, Maryland
Salon I
Congressional Ball Room

Tuesday, January 13, 1998 - 7:00 p.m. - 9:30 p.m.

Reports:

Resource Planning Workshop----------Lisa Brooks
Sequencing PI Meeting----------Jane Peterson
Function Workshop----------Elise Feingold

Other planned workshops:

Mouse----------Bettie Graham
Informatics/databases----------Lisa Brooks

Other items of interest:

DOE planning activities----------Marv Frazier
NIGMS workshops----------Chuck Langley, Lee Hartwell

Update on FY 1999 budget----------Elke Jordan

Wednesday, January 14, 1998 - 8:30 a.m. - 4:00 p.m.

Discussion of outline of 5-Year plan

Leroy Walters will join us around lunch time to report on ERPEG activities

*****We will work through lunch—food will be provided.**

By the end of the day, we should have a pretty clear idea of what the goals will look like for the areas where workshops have already taken place. The next 2-3 months will then be spent drafting the actual document, refining the goals, filling in the gaps, etc. No other meeting of the subcommittee is scheduled until May. We need to decide whether this is sufficient, i.e., can we work by e-mail and possibly conference call in the interim?

At the May 5/6 meeting we will need to integrate the DOE and NIH aspects into a draft that can be presented to the community at Airlie House on May 28/29.

Workshop on Human DNA Sequence Variation
Report

The National Human Genome Research Institute (NHGRI) convened a Workshop on Human DNA Sequence Variation on March 31 and April 1, 1997, on the campus of the NIH, to address scientific issues concerning how the reference human DNA sequence might be annotated by information on DNA sequence variation. This was the initial workshop in a series of planning meetings that will be organized over the next year by the NHGRI to help the Institute identify the most important scientific questions for future investigation and investment in anticipation of the completion of the first reference human DNA sequence by the year 2005, and the need to consider making investments now to maximize the beneficial consequences of genome research.

The agenda for the workshop, the questions discussed, and a list of participants, is attached. In summary, the workshop discussions led to the following conclusions:

1. There is a critical and immediate need for NIH to stimulate and support pilot projects investigating a number of important questions in population genetics. Research needs include quantifying DNA variation, understanding how it varies across the genome, and how DNA variation (deleterious, advantageous and neutral) arises and is maintained in human populations.

2. A defined resource of cell lines and/or DNA, that would be appropriate for studying a variety of questions pertaining to human sequence variation, and would be generally available to the scientific community, would potentially be of very great value. There are a number of important scientific and ethical questions that must be addressed before such a resource is developed, and further discussion of this possibility is enthusiastically encouraged.

3. Further research into efficient methods of detecting DNA sequence variation and of genotyping, particularly research aimed at increasing the sample throughput and decreasing the cost of analysis, is necessary.

As progress is made toward the determination of the complete genomic DNA sequences of the human and of several non-human organisms, consideration about ways in which to augment information about the genome sequence is increasing. Although sequence data can be annotated in several ways, two obvious areas of interest are determination of sequence variation and the effect of such variation on functions encoded within the genome.

Human genetics is critically concerned with how variation in gene sequences is related to variation in the function of genes and gene products. Neutral variation, which does not affect gene function, provides information on human population structure and the history of chromosomes. Thus, the identification, classification, quantification

and analysis of sequence variation is expected to constitute one of the most powerful, and direct, approaches to the study of a wide range of important biological questions. The reference genome sequence will provide scientists with the basis for measuring sequence variation, assessing how variation in specific genes is associated with complex phenotypes (and common diseases), how sequence variation affects gene function and biochemical pathways, and how human genetic variation has been shaped by biological processes of natural selection and evolution. At present, there are neither significant data on the nature and extent of DNA sequence variation in the human genome, nor much discussion on the gamut of biological studies that could benefit from broad knowledge of genomic variation. Thus, the workshop was organized to discuss three primary scientific issues: (1) the kinds of research that require or benefit from information about genetic variation, (2) the characterization of DNA sequence variation in the human, and (3) the technologies necessary for determining, assaying, and interpreting variation across the genome.

The study of variation in human genes, either through the analysis of phenotypes known to have a genetic basis or through known gene products (blood groups, serum proteins, enzymes), has a long history. This has led to our current understanding of the nature of genetic variation in humans. However, there are many inherent problems in such classical studies, including: studies have largely involved genes already known to be variable in human (often European) populations, only a small number of genes has been studied and only some types of genomic changes have been monitored. Moreover, as these studies were largely concerned with variations that affect protein composition, they cannot accurately reflect the degree and nature of genetic variation at the DNA sequence level.

The development of DNA sequencing technology and the initiation of large-scale DNA sequencing projects has now led to the ability to directly measure variation in genomic DNA. This means that the entire genome, and not simply the recognized coding sequences, is accessible for analysis. The initial results of genomic sequence analysis indicate that, on average, there is one variant nucleotide (nt) per 1000 nt (one kilobase, or kb) screened, confirming the results of many studies performed in the 1980's using RFLPs. For example, one laboratory found one single nucleotide polymorphism (SNP) per 973 nt (comparing approximately 300 kb in three individuals by one method; similar numbers came from a study comparing one megabase in eight individuals by a different technique). In this study, the evidence for clustering of variation was, at most, slight. However, another laboratory reported considerably more difference in the local frequency of variation, which ranged from one difference per 860 nt to as little as one difference per 10 kb. As more human DNA sequence is determined, and as methods for resequencing the same region from several individuals and populations are improved, more data of this type will be collected.

Much of the discussion at the workshop focused on the use of sequence variation information in the identification of the loci underlying multigenic traits, in particular, complex diseases. Specifically, the usefulness of generating a high resolution ("third

generation") map of sequence variation at single nucleotide positions for the identification of the genes underlying such traits was a topic of primary interest. The discussion of this subject focused on questions such as the marker density necessary for the maps to be useful in such studies, the types of human populations that should be studied for complex disease mapping, the extent of haplotype information, and the types of family structures that might be used. The issue of marker density is unresolved and requires additional theoretical study.

An important issue in complex disease genetics, about which little is currently known, is the nature of genetic variation which underlies such phenotypes. Are these diseases like rare mendelian phenotypes, in that the multiple loci harbor rare and generally new mutations of large to intermediate allelic effects? Or, are they due to a combination of common genetic variants each of which has a small allelic effect ? If the latter alternative is true, then it would be advantageous to screen the human population and catalogue common variants (as has been previously done for the HLA system); however this will not be a practical approach until information about full-length coding sequences for many genes is available.

Another unanswered question relates to whether common variants, known to be susceptibility factors in some complex diseases, are likely to be recurrent or have one or a few origins. If gene variants have few origins, then for each variant a considerable segment of DNA surrounding the variant allele (haplotype) will be shared between individuals harboring the variant. For gene mapping, then, the question of the required marker density centers on the size and ability to recognize these haplotypes or "ancestral blocks" (contiguous regions of DNA that have largely been inherited without recombination in human evolution). As the functional information is contained within these blocks, association studies can be used to correlate haplotypes with specific phenotypes. It was suggested that approximately 10-20 single nucleotide polymorphisms (SNPs) per block would be sufficient for characterizing the human genome. This could translate to a map of 30-60,000 markers to analyze blocks of a megabase in size, which might be the case in studying a unique population with a recent ancestry. More typically, for an outbred population (with a lot of mixing and old mutations), the blocks will be considerably smaller, requiring a comparably larger number of makers. Block size is critically dependent on the population under study. Such dense SNP maps would enable the study of diseases from appropriate patient samples by association, without the necessity of family samples. This would reduce the cost of disease studies and, more importantly, genetic studies could be better designed with respect to the phenotype.

Human disease studies are best performed in populations in which genetic heterogeneity, both locus and allelic heterogeneity, can be minimized, and populations that satisfy these criteria need to be identified. Unfortunately, these characteristics for the specific disease loci cannot be readily measured. In general, culturally and geographically isolated populations satisfy this rule, but often do not have a sufficient number of cases for disease mapping to proceed with precision. Better methods and better parameters for characterizing human populations are necessary. As at other meetings, the issue of

complex disease mapping and gene identification generated many opinions, but it was clear that there is not enough information available to answer the crucial questions conclusively. Thus, there is a critical need for NIH to stimulate and support pilot projects investigating a number of important issues, including how to make SNP maps, how to survey common sequence variants (how do we detect all common variants? is there benefit to restricting analysis to only those in coding sequences or regulatory regions?), how to explore the power of different "populations" to identify disease genes (what is the effect of a population's size and age on its usefulness in detecting disease genes?), how to survey ancestral haplotypes (what specific populations should be used in such studies?), and how to develop a bioassay for population heterogeneity.

A second idea that emerged from the workshop was the potential usefulness of a reference set of samples that could be used by scientists with diverse research interests as a resource to characterize and study human sequence variation. It was suggested that a collection of 500 trios (i.e., sampling both parents and one child) comprised of a relatively small number of "groups" (five) constructed in such a way that it could properly "represent" the U.S. population (if this were to be developed by a U.S. funding agency) would be extremely valuable. It was recognized that there are a number of scientific questions that need to be considered in developing such a resource and, beyond the scientific questions, there are very important ELSI issues that must also be addressed in the construction of such a collection. Further discussion of this concept is clearly needed.

The workshop discussed a variety of technologies that currently exist and which can be applied to genetic variation studies on a genome-wide scale. These methods, in their current implementation, are efficient either for the detection of genetic variation or for assaying specific variants in multiple samples, but not for both purposes. Further research into efficient detection and genotyping methods, particularly research aimed at increasing the sample throughput and decreasing the cost, is critically necessary.

A Resource for Discovering Human DNA Polymorphisms

Goal: A resource of DNA samples and cell lines that can be used to discover polymorphisms in the U.S. population.

Samples: The resource will include cell lines and DNA from 500 unrelated individuals, female and male. In addition to the complete collection, there may be one or more predefined subsets, encompassing the same range of diversity as the complete set. This subset is for use in technology development efforts, which require a diverse set of samples but do not aim to detect variation in the complete set.

Geographic origin: To maximize the chances of discovering the common DNA sequence polymorphisms, the individuals sampled will be U.S. residents who have ancestors from the major geographic regions - Africa, Asia, Europe, and the Americas. Individuals with ancestors from more than one region will be included.

Anonymity and lack of information on individuals: Information about geographic origin and sex will be collected in order to ensure a diverse sample. Summary information describing the complete collection and any predefined subsets as wholes will be made available, but no identifiers will be associated with the individual samples. All identifying and phenotypic information will be removed from the individual samples so that the links to the individual donors will be irreversibly broken.

Accessibility: The material in the resource will be available to any investigator, provided that the proposed use of the material has been reviewed by an IRB and approved or designated as exempt.

Timing: The repository should be ready to distribute material by October 1998.

Informed consent: All samples will come from individuals who have provided informed consent to be part of this resource. The informed consent material will explain that the information collected through the use of this resource will be used for a wide range of genetic studies that will address many questions, some currently unknown.

Source of samples: The cell lines will come from more than one source:

- the CDC NHANES III study, with informed consent obtained for this study;
- existing collections, when informed consent for this study can be obtained; or
- on-going collection studies, with informed consent obtained for this study.

Location: The Coriell Institute seems to be the most appropriate repository.

Database: A central database will receive information on all variants found. Since all identifying information will have been stripped off, each sample will be identified only by a unique sample number. The alleles detected for each polymorphism for each sample will be recorded in this central database.

NHGRI WORKSHOP ON--RESOURCES FOR DETECTING GENETIC VARIATIONS
DECEMBER 8-9, 1997
PARTICIPANTS LIST

Aravinda Chakravarti
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106
[redacted]

Linda Burhansstipanov
Director of AMC
Cancer Research Center
1600 Pierce Street
Denver, CO 80214
[redacted]

Kenneth Buetow
Fox Chase Cancer Center
7701 Burholme Avenue
Philadelphia, PA 19111-2412
[redacted]

Georgia Dunston
Professor and Interim Chair
Dept of Microbiology
College of Medicine
Howard University
520 W. Street NW
Washington, DC 20059
[redacted]

Jonathan Friedlaender
Professor of Biological and Anthropology
Dept of Anthropology
Gladfelter Hall, Rm 215
Temple University
Broad Street, Berks Mall
Philadelphia, PA 19122
[redacted]

Bronya Keats
Dept Biometry and Genetics
Louisiana State University Medical Center
1901 Perdido Street
[redacted]
Email: biombjk@lsumc.edu

Charles Langley
University of California
Davis Center for Population Biology and
Section of Evolution and Ecology
Davis, CA 95616
[redacted]

D. Andrew Merriwether
University of Michigan
535 W. William St.
1045 NUBS
Ann Arbor, MI 48109
[redacted]

John Moore
Professor and Chair
Dept of Anthropology
Box 117305
University of Florida
Gainesville, FL 32611
[redacted]

Robert Nussbaum
Chief of Laboratory Disease Research
National Human Genome Research
Institute, Bldg 49, Room 4A72
National Institutes of Health
49 Convent Drive
Bethesda, MD 20892
[redacted]

Madison Powers
Senior Research Scholar
Kennedy Institute of Ethics
Georgetown University
1437 37th Street NW
Washington, DC 20057
[redacted]

NHGRI WORKSHOP ON--RESOURCES FOR DETECTING GENETIC VARIATIONS
DECEMBER 8-9, 1997
PARTICIPANTS LIST

Nancy Press
Associate Professor
Dept of Psychiatry and Biobehavioral Sciences
School of Medicine UCLA
760 Westwood Plaza
Los Angeles, CA 90024
[redacted]

Edward J. Sondik
Director of National Center for Health Statistics,
Centers for Disease Control and Prevention
Presidential Bldg, Rm 1140
6525 Belcrest Rd.
[redacted]

Karen Steinberg
Chief of Molecular Biology Branch
National Center for Environmental Health,
Centers for Disease Control and Prevention
4770 Buford Highway NE
Mailstop F-24
[redacted]

Diane Wagener
Acting Director of the Division of Health Promotion Statistics
National Center for Health Statistics,
Centers for Disease Control and Prevention
Presidential Bldg, Rm 770
6525 Belcrest Rd.
[redacted]

LeRoy Walters
Director
Kennedy Institute of Ethics
Poulton Hall, Rm 222
Georgetown University
1437 37th St. NW
Washington, DC 20057
[redacted]

Bruce Weir
Professor of Statistics and Genetics
Dept of Statistics
Patterson Hall Rm 220
Box 8203
North Carolina State University
[redacted]

Kenneth Weiss
Distinguished Professor
Dept of Anthropology/Biology
409 Carpenter
Pennsylvania State University
University Park, PA 16802-3404
[redacted]

DRAFT 01/07/98

REPORT ON THE PLANNING MEETING OF THE NHGRI LARGE-SCALE SEQUENCING PRINCIPAL INVESTIGATORS, DECEMBER 18, 1997

SUMMARY

The Principal Investigators of the NHGRI-funded large-scale sequencing centers met with members of the Council Planning Subcommittee and NHGRI staff to discuss the critical issues in large scale sequencing that should be taken into account in formulating the next NHGRI five-year plan. These issues included: 1) sequencing costs; 2) scale-up plans and sequencing capacity; 3) technology development; 4) the merits of centralized mapping; 5) the role of mouse sequencing; and 6) any other issues the PI's felt to be critical. The major points made by the Principal Investigators were as follows:

1. On the basis of progress made in the past few years, the PI's are optimistic that the human genome can be sequenced by 2005.

2. Current unit sequencing costs are about $0.50 per base pair. However, concern was expressed about the likelihood of continuing decreases in sequencing costs. Several PI's stated that it is not clear how the next 2-fold cost reduction, which NHGRI projections have assumed, will be achieved.

3. Cost accounting is difficult, but uniformity is desirable. It is apparently widely perceived that NHGRI currently focuses on a 'dollars in, bases out' calculation, which the PI's think is too simplistic [staff note: this is not actually a complete description of the NHGRI approach, which does begin with "dollars in, bases out" but then goes on to ask the sequencers to back out (and therefore identify) each of those costs which are not appropriate to consider as production costs]. It was suggested that accounting procedures that break down costs according to process (for example, with 'per-lane' accounting) may be more useful for assessing which parts of the process are amenable to cost decreases in each center, over the longer term. 'Per-lane' accounting may also be more useful for reviewers to directly compare performance of centers.

4. The $60M per year that NHGRI has proposed to devote to large-scale human sequence production may not be adequate. It was argued that holding spending to this level would discourage groups from exploring more efficient ways of sequencing in the long run and does not even account for the full cost of scale-up, e.g., it does not accommodate the cost of new facilities. Some argued that, in the short term, costs may actually increase rather than decrease, particularly to maintain good quality sequence. The NHGRI cost model was also criticized for not allowing for (1) a large initial investment that may result in long range cost reductions and (2) large failures. The suggestion was made to reserve more (as much as all) of the NHGRI budget for large scale sequencing and only funding other parts of the programs from the savings achieved by reducing sequencing costs.

5. Maintaining high sequence quality is challenging (and increases costs), but is critically important to the success of the project.

6. A number of local problems common to several centers were identified: (1) space is limiting capacity; (2) retention and training of good personnel is difficult, and (3) attracting talent to the field is difficult. The PI's suggested that these problems could be ameliorated if NHGRI were to adopt ambitious goals, beyond the human genomic sequence. This would help them convince university officials, good personnel, and outside talent that there was a long-term future for the sequencing centers.

7. Only one center declared plans to reach a capacity of 100 Mb per year. Most of the others reported plans to scale up to 20–50 Mb per year (including some groups, which had been thought to have interest and potential to scale to 100 Mb per year).

8. There was discussion of the concept of supporting sequence-ready map production for both the human and mouse genome separately from sequence production. However, no consensus developed that the current NHGRI policy of directly linking mapping and sequencing should be changed.

9. Technology development is sometimes difficult in an academic environment but private industry is dissuaded from it by a perception of lack of profitability. Ambitions goals beyond completion of the human sequence may help increase enthusiasm for technology development.

10 . Sequencing the mouse genome should be taken to be a goal of the NHGRI and should be aggressively pursued once the infrastructure for sequencing the human genome is fully funded.

PLANNING MEETING REPORT

The NHGRI-supported investigators engaged in large-scale sequencing (PI's) met on December 18 and 19, 1997. The first day of the meeting was to provide the PI's an opportunity to discuss with NHGRI staff and the Council Planning Subcommittee prospects for achieving the sequencing goals of the Human Genome Program (referred to herein as the 'planning meeting'); this portion of the meeting was open to the public and is summarized below. On the second day of the meeting (referred to as 'the PI's meeting') which was closed to the public, participants discussed other common issues outside of those related to the five year planning effort.

In preparation for the planning meeting, the subcommittee sent the participants a set of questions (attached). Each PI was asked to prepare a ten minute presentation addressing these questions and other critical issues that they believe must be addressed for the HGP to succeed in achieving the sequencing goals of the project. These presentations, and the resulting discussion will be used by the subcommittee in writing the next NHGRI five-year plan.

Drs. Collins and Chakravarti opened the planning meeting with an overview of the planning process, and asked the participants to address the genomic sequencing goals and milestones that should be achievable in the next five years. Dr. Chakravarti charged the participants to discuss the range of questions sent to the PI's in advance of the meeting by the planning subcommittee. What are the critical issues that the subcommittee must consider in planning to finish the human sequence? What are the costs? How will sequencing capacity have to change? What will the infrastructural needs be? Dr. Chakravarti encouraged the PI's to contact the subcommittee members if they had any comments on these issues, beyond those in the present meeting.

Discussion: Cost model

To start the discussion, Dr. Guyer presented the NHGRI cost model on which current plans have been made. This model assumes that:

- NHGRI will fund 60% of the human sequencing effort
- Costs will decrease with a half-life of 4 years, with a starting point of $0.50 per base in 1998
- NHGRI will make an annual investment of $60 million (adjusted for inflation)

Other identified commitments outside the amount now intended for sequencing activities include SNP's, technology development, databases, model organisms, some functional analysis, training, and ELSI efforts. NHGRI staff projects that, after meeting these commitments, $12–20 M in additional funding will be available for new spending after 1998 (assuming a 3% increase in budget). Dr. Collins felt there was good reason to be optimistic that the NHGRI budget will increase.

Reports from Principal Investigators: critical issues that need to be addressed in large-scale sequencing

Each PI addressed the questions posed by the subcommittee in a brief presentation

1. Washington University/Robert Waterston. Dr. Waterston summarized the successes of the past few years: demonstration that the clone supply for sequencing the human genome will be adequate and that contigs can be built with those clones; demonstration that the shotgun strategy works over megabase-sized genomic regions; and EST projects and new software programs have begun to demonstrate that we will be able to interpret the data. The two remaining questions are whether enough centers can scale up to complete the sequencing goals on time and whether the necessary cost reductions can be achieved. He estimated that his center will be able to scale to 120 Mb per year in three years, but he is uncertain how the next two-fold reduction of costs will be achieved. He estimates that the current cost in his center is ~$0.50 per base.

   Dr. Waterston commented that mouse sequence would add value to the interpretation of the sequence. He also believes that the HGP faces significant challenges in the form of distractions and competition. Specifically, retaining good personnel is increasingly difficult, since they get attractive offers to work on other projects. In his opinion, a $60 million expenditure by NHGRI is too small an amount to represent enough of a commitment to keep talented people interested. The difficulty of keeping good personnel is complicated by frequent funding decisions, i.e. lack of long-term commitment to a center. Dr. Waterston suggested that this could be partially addressed by NHGRI planning on continued increases in sequencing capacity beyond 600 Mb per year. This would encourage talented individuals to the project.

2. University of Oklahoma/Bruce Roe. Dr. Roe plans to scale up to 20–40 Mb per year over the next five years, by following the example provided by the Washington University group. The greatest current bottleneck in his center is finishing, and some technology development in this area would be useful. The major limitation to longer-term scale-up at the University of Oklahoma is space. Dr. Roe remarked that annotation, which he stated is cut from sequencing grant applications, should not be the responsibility of the PI's, but rather of the NLM informatics infrastructure. Further, NLM should put confidence values in the sequence database. He agreed with Dr. Waterston that $60 Million per year would be too small an investment by NHGRI and that it is difficult to keep staff. He cautioned NHGRI to limit its investment in new technology that will not be implemented in sequencing labs by 2005, although he acknowledged the need for new technology to drive the costs down.

3. University of Washington per Maynard Olson. Dr. Olson made several points: 1) NHGRI should not waffle on quality; 2) NHGRI should be realistic about costs; 3) planning and resource allocations must be based on past performance and quality; 4) NHGRI should develop appropriate performance assessment mechanisms (in his opinion, the current "$ in, base pairs out" approach is flawed (see below)), and 5) NHGRI's projected $60 million investment is not enough. He predicted that there will be substantial failures in the program and the proposed project costs do not allow for such failures. He argued that we do not yet know how to "turn dollars into high quality bases." In a separate presentation, Dr. Olson reviewed his center's approach of assessing costs on a 'per-lane' basis. He stated that this gives them a better idea about how much each part of the process costs, and he cautioned that a savings in one step (e.g., as provided by a higher capacity sequencer) does not necessarily lead to an overall increase in efficiency; if care is not taken, such 'improvements' can be more than canceled out by the increases in cost it may cause in another step (e.g., higher finishing

costs if run length or quality was reduced). Per-lane accounting may be useful for comparisons between groups. Equipment costs also need to be realistically amortized.

Dr. Olson argued that mouse sequencing should be undertaken soon because of its value in interpreting the human genome. He also emphasized the need to keep enough viable centers actively engaged in large-scale sequencing to ensure that we have the capacity to complete the project

4. Stanford University/Richard Myers. Dr. Myers plans to scale up his center to 30 Mb per year in two years. The current estimated cost of sequencing in his center is $0.64 per bp, but he believes this will decrease to $0.15–0.25 per bp. He urged NHGRI to maintain quality and to increase the minimum contig size used for counting sequencing progress to greater than 30 kb. He believes that the most critical issue is management and he is concerned that there may not be enough sequencing groups involved to accomplish the goals.

5. Whitehead Institute/Eric Lander. Dr. Lander made three points. First, the current NHGRI goal is "faint-hearted" and those involved in human genome sequencing are being distracted by other projects . He argued that more exciting goals will attract good talent and more resources (an example of an ambitious goal is "one mammalian genome a year"). Second, the NHGRI cost model may be incorrect in assuming a logarithmic decrease in costs. He pointed out that for many industrial processes costs, an "S"-shaped curve is a better description of the cost curve. He also argued that cost decreases are driven by investment, and that current funds are inadequate to stimulate such cost reductions. Third, Dr. Lander discussed the benefits of funding separate mapping centers to provide sequence-ready maps to the sequencers. He reasoned that the autonomy of such centers would allow them to innovate and economize in a way that is not currently possible under the existing model where mapping is an integral part of a sequencing effort.

6. Baylor College of Medicine/Richard Gibbs. Dr. Gibbs estimated the cost per base at his center to be $0.50 for the last year, and expressed a desire to scale to 22 Mb per year by 1999. He described his group's experience in doubling the efficiency of most parts of the sequencing process they have analyzed (for example, reactions per month) over several successive six-month periods, without an increase in personnel. However, he noted that these increases in productivity in component processes are not yet reflected in the overall costs. and the quality of the sequence suffered as the center scaled. This increase in efficiency was achieved with little contribution from automation, which Dr. Gibbs expects will contribute to future increases in productivity. He urged NHGRI to provide funds for importing technology.

7. University of Texas Southwestern Medical Center/Glen Evans. Dr. Evans' group is finishing sequence at a rate of 1 Mb per month. He noted that the increases in efficiency have resulted from management changes. The center's goal is to produce 12 Mb of finished sequence by August 1998 at the current funding level. With increased funding, he expects to produce 30–45 Mb in the following year. Over the last ten months, sequencing costs in his center have varied widely ($2.00—~$0.20-0.40 per base) due to a variety of factors, such as closing gaps in clones containing GC-rich repeats (about 1 in 3). 20 % of the center's funds have been spent in developing and exporting new technology. Dr. Evans asserted that automation/robotics should not be expected to affect efficiency in centers with a sequencing capacity in the 2–10 Mb per year range.

8. The Institute for Genomic Research/Mark Adams. Dr. Adams focused on issues related to the scale of sequencing, rather than its cost. He pointed out that there are both efficiencies and inefficiencies of scale. Sequencing costs in his center are currently ~$0.50 per base pair. The center had planned a 4-fold increase in production of human sequence for the current year, but because of difficulties with the production of sequence-ready maps, the projected scale-up has not occurred. TIGR's current sequencing capacity for all sequencing efforts is 20 Mb per year and Dr. Adams projected that it could increase to 50 Mb per year. The TIGR center is experiencing problems with high personnel turnover, and has instituted a separate training program to deal with the need to continually train new personnel. Dr. Adams hopes to increase the automation within the center to decrease dependence upon personnel. He felt that separate mapping centers were problematic, as individual sequencers have different needs depending on strategy and map quality is difficult to assess.

9. University of California, Berkeley/Gerald Rubin. Dr. Bruce Kimmel presented plans to scale up the Drosophila sequencing effort to 30 Mb per year (the current rate is 1.1–1.7 Mb per month, at a cost of ~$0.47). He believes that the cost can be reduced to $0.30–0.35 per base pair as a result of efficiencies of scale without further automation. Dr. Kimmel broke down his center's costs by process (mapping, administrative, etc.). He cited the help in industrial engineering his center received from Motorola, and feels that a plateau in efficiency and cost reduction will be reached without further technology development. However, he believes that careful analysis of where bottlenecks occur is needed. Dr. Rubin commented that the center's output is currently limited by space.

Dr Chakravarti and other attending members of the planning subcommittee asked participants for further discussion of several specific questions , emphasizing that this was for planning, and not review purposes. Will we reach a cost of $0.25 in four years? How are we building capacity? At what rate can we invest new funds without wasting money?

Many participants reiterated that it was difficult to foresee how costs would decrease much beyond the current cost of ~0.$50 per base pair. Others felt that that scale-up will actually lead to temporary cost increases, unless costs (e.g., expensive equipment) are amortized over a several years. It was suggested that NHGRI should assume a 'worst-case' scenario and reserve the total NHGRI budget for large-scale sequencing, only funding other programs from any economies achieved.

Participants re-emphasized the importance of setting ambitious goals to attract talent, maintain staff and generate enthusiasm among decision makers. Participants suggested expressing the goal in terms of capacity (e.g., 1 gigabase per year in six years at a certain cost) as well as in terms of simply finishing the human genome. One participant re-emphasized that quality considerations must be included in stating any goals, otherwise the costs of the most difficult parts of the process will be deferred.

Because the planning subcommittee needs a clear picture of what each center can do, Dr. Chakravarti asked the PI's for more input on this issue, either in discussion at the PI's meeting or by e-mail. Each PI was asked to provide Dr. Chakravarti with an analysis of cost for the upcoming planning committee meeting.

Discussion: Should mapping efforts be centralized?

Several years ago, NHGRI developed a policy that its funding of efforts to construct sequence ready maps must be partnered with a sequencing group that will sequence the mapped clones. This policy was based on recommendations from NHGRI-supported investigators who were funded to carry out large scale sequencing. Recently however, several investigators have questioned whether this policy should be reexamined in light of progress in mapping and the difficulty encountered in some centers in maintaining a high quality, high throughput and low cost mapping and sequencing effort.

The PI's at the meeting were divided on whether it was either technically feasible, practical, or desirable to make such a change in funding sequence ready mapping. The arguments for funding dedicated mapping centers were that such centers could better monitor cost and quality and would develop a community that would encourage advancement in mapping techniques. The arguments against separately funded sequence-ready mapping were that just-in-time mapping as practiced in several centers is working well, mapping quality standards are hard to agree upon and implement and mapping should serve the needs of a sequencing group which in turn will drive the mapping effort to produce quality maps that are contiguous, centers focussed on mapping large regions of the genome will be distracted from the important goal of closing gaps in small but frequently occurring unclonable regions, differences in sequencing strategy will render centrally produced maps less useful to some investigators, and finally there will be additional informatics needs to deal with the dissemination of centrally constructed maps.

There was some agreement that centralized mapping may be more practical for the mouse than for human, since human mapping already has an infrastructure. Further, there is less liability due to changing ethical considerations with a mouse map, if it is made substantially in advance of sequencing. Some participants suggested that NHGRI consider a pilot study to determine whether useable clone maps could be constructed centrally, with adequate quality controls and contiguity standards (perhaps 25 kb resolution over 10 Mb contiguity).

A proposal was made, and debated, for increasing the sequence contiguity requirements for NHGRI finished sequence to 500 kb (from its current 30 kb) in March, 1998.

Discussion: Technology development

The participants were asked to address the following questions: What technological improvements are needed over the next 5 years to successfully scale-up? What is needed in the longer term? Are current mechanisms to implement new technology adequate?

Participants raised a variety of issues. To the extent that technology can replace people, it is desirable, since automation can offset training and other personnel costs. However, there is a loss of flexibility with a large commitment to automation. Technology development is very expensive, and generally poorly suited to an academic environment. However, it is also difficult to interest

private companies in developing technology for genomic sequencing with the type of throughput needed for large-scale sequencing because of the perception that there is limited profitability. It is even difficult to interest talented engineers and physicists to participate in the project, because of the perception that the genome project is finite. One participant encouraged NHGRI to support technology development separately from large-scale sequencing, because currently the 'dollars in, base pairs out' model discourages inclusion of technology development if it does not decrease sequencing costs within a short time. Another participant felt that the main barrier to implementation of new technology was the ability to identify good new technology operating in other labs.

One important consensus emerged from this discussion: participants were in general agreement that there needs to be more cross-fertilization between the sequencing centers about technology development and implementation, as well as other issues related to technology. Participants urged the creation and funding of 'working groups' and other less formal mechanisms to explore specific problems and report findings at PI meetings. Topics of interest for working groups include: What is the importance of read length in increasing efficiency and decreasing cost? What is the best way to do cost accounting? How do we facilitate technology transfer? Are there ways to ensure better representation of repeat-rich, or other 'unclonable' regions of the genome?

Most of the discussion centered on "production-related" technology development, and how to implement it in the centers. There was a lot of enthusiasm for creating 'working groups' and other less-formal means of encouraging cross-fertilization in this area. It was also reiterated that the perception of the NHGRI goals as being more ambitious was an important factor in attracting talent and technical infrastructure.

Discussion: Mouse sequencing

The current 5-year plan includes production of a genetic map and sequencing of syntenic regions of the mouse genome as goals. There is general enthusiasm for including sequencing of the mouse genome as a long term goal in the next 5-year plan, but how and on what timetable? The participants were asked to share their opinions on this topic with the planning committee.
The value of sequencing the mouse genome in order to interpret the human genome was repeatedly stated. The participants urged NHGRI to declare its intent to sequence the mouse genome. The discussion centered on how to prioritize mouse compared to human sequencing, and also on whether it was worthwhile to construct a sequence-ready map in advance of the genomic sequencing.

Several participants cautioned that, while it is important to sequence the mouse, it is also important to ensure that the capacity is available to complete the human sequence first and that mouse sequencing should be limited until then. One participant pointed out that with the human sequence in hand, it would then be reasonable to focus on the most informative regions of the mouse genome based on biological interest (e.g., Hox clusters).

Participants were again divided on the merits of constructing a sequence-ready map in advance of the sequencing. Some argued that it was too risky to tie sequencing to a single clone resource

constructed with current technology. Others thought that a clone-based map was desirable, and could be constructed with clearly defined quality standards, and suggested a pilot project to see if this was feasible.

Some of the genomic goals of the mouse community were also discussed. Participants pointed out that even if a complete, ordered BAC map were feasible, it was not useful for positional cloning.

October 22, 1997

TO: PI's of NHGRI Pilot Project Sequencing Centers

FROM: Five Year Scientific Planning Subcommittee

RE: Sequencing goals for completion of the reference human genome sequence (RHGS)

As a part of our charge to help develop the next five year plan for the NHGRI we, the Scientific Planning Subcommittee, want to follow up on our intention to involve the large-scale sequencers in the planning process and get your input into our considerations. We recognize that the centerpiece of the plan will have to be the completion of the first reference human genome sequence (RHGS). Consequently, we want to evaluate the current status of DNA sequencing prior to making our recommendations for the sequencing plans and goals for the next five years.

The NHGRI's current plans to continue funding sequence production (excluding non-production-related technology development to be funded separately) at $60M per year are based on the assumption that the cost of sequencing will drop from a current average value of $ 0.50/base in 1998 with a half-life of 4 years. If these assumptions are correct, this will allow the NHGRI goal of finishing 1.75 Gb finished sequence to be met by 2005. The Planning Subcommittee believes that it is important to critically evaluate our current and projected production capability, to compare this estimated capability with our goals, and to examine the assumptions being made to ensure that the resources, goals and abilities are mutually compatible. Therefore, we are requesting your input, as the people who have the most experience in this endeavor, and information about your current and projected capabilities and your ideas about the future, so that the Subcommittee's discussions will be more informed. We will meet with you on December 18, at the PI meeting, and would like those discussions to be as substantive as possible. We think that the issues and questions outlined below will provide a useful basis for discussion, so we are requesting that you come to the meeting prepared to address these questions, and any other information or approaches that you think will be useful to us in our deliberations.

We recognize that some of these questions may be difficult to answer in as much detail as we might like to have. We request that you provide us with a good faith estimate in these circumstances and explain your

assumptions to us. It is critically important that the answers we come up with be as reliable as possible. This is a critical time in setting realistic goals and effective scientific and funding plans, so that we do not have bigger problems downstream. We appreciate your cooperation and interest and look forward to your candid responses.

1. (a) We need to generate an accurate estimate of current costs for producing finished sequence according to current standards on a "$ in/finished base out" basis. This should be calculated to include production-related technology. The estimated costs should be fully loaded in that overhead should be included. Define as clearly as possible how these numbers are generated.

(b) Can you make an estimate of how your costs and effort distribute between completing "difficult" versus "less-difficult" sequences?

(c) How much of the effort (time and $'s) is at clone characterization and mapping, sequence generation, assembly, and finishing?

2. (a) Using your current proven production sequencing technologies and approaches, what level of ramp-up in sequencing effort is feasible at your location and how long would it take to accomplish this scale-up. The answer to this question should cover the next five year planning period (1998-2003). Estimates of one-time startup costs for such fixed items as building renovation, equipment, office space etc., that would have to be paid up-front, but are not a part of the continuing sequencing costs, should be estimated separately.

(b) Will the cost per base be about the same, higher or lower after these ramp-ups in production? Why? Here we would like to know how your sequencing operation is organized and how it scales, or is expected to scale, with size. What are the current bottlenecks and what happens as the scale of your operation is increased?

3. (a) Using your best estimate of possible technological improvements in your production sequencing approach, what level of ramp-up in sequencing effort is feasible at your location and how long would it take to accomplish this scale up ? The answer to this question should cover the next five year planning period (1998-2003). Estimates of one-time startup costs for such fixed items as building renovation, equipment, office space etc., that would have to be paid up front but are not a part of the continuing sequencing costs, should be estimated separately. Assuming these improvements and ramp-ups are feasible, what is your best estimate of the sequencing cost that you can achieve? What fraction of the future costs will be for mapping and what fraction for sequencing?

(b) How much of this improvement in production rate and costs will depend on the adoption/adaptation of available proven methods to in-house production?

(c) How much of this improvement in production and cost will depend on the development of methods yet to be proven in production sequencing?

4. (a) Do you foresee the development of a fully integrated system based on current technologies or rapid evolution of methods. Why and how?

(b) Can you support the assumptions of cost decrease the NHGRI has projected and if so on what basis?

(c) Are the current plans for making reagents suitable for sequencing sufficient? If not, what remains to be done?

(d) We would like your feedback on what you believe are the most critical questions that will require a solution in the next 2 years (1998-2000) and in the period 2000-2005? What are your concerns and what can the NHGRI do to help?

MEETING OF NHGRI-SUPPORTED INVESTIGATORS ENGAGED IN LARGE-SCALE SEQUENCING

HYATT REGENCY BETHESDA
DECEMBER 18, 1997

8:30 AM	**Introductory Remarks**	**Francis Collins** **Aravinda Chakravarti**
8:45 AM	**Budget Overview and Review of NHGRI Cost Model**	**Mark Guyer**
9:00 AM	**P.I. Statements on Critical Issues in NHGRI Sequencing Plans. Each PI will have ten minutes to discuss what he considers to be the most critical issue(s) that need to be addressed concerning NHGRI's plan for completing the human DNA sequence. This includes issues raised in the list of questions sent by the planning committee, the current status of human genome sequencing, the current NHGRI cost model (current and future projections for the sequencing budget), feasibility of plans for scale up, or other issues.**	**Bob Waterston** **Bruce Roe** **Maynard Olson** **Rick Myers** **Eric Lander** **Richard Gibbs** **Glen Evans** **Mark Adams** **Gerry Rubin**
10:30 AM	**Coffee Break**	
10:45 AM	**General Discussion -- Is the NHGRI cost model for completing the human genome sequence reasonable? If not, what is a better approach to estimating the requirements for sequencing the human genome?**	
12:30 PM	**Lunch**	
1:30 PM	**Mapping: Is "just-in-time" mapping a workable strategy for successfully sequencing the genome or should we consider other models, such as one (or more) centralized mapping effort(s)?**	
2:30 PM	**Technology Development: What improvements are needed over the next 5 years to successfully scale up the sequencing effort to meet the projected targets? What improvements are needed beyond that time period to complete the project? Are current mechanisms for implementing new technologies into sequencing centers adequate?**	
3:30 PM	**Mouse Sequencing: We are assuming the mouse genome will be sequenced. How and on what timetable should it be completed?**	
4:30 PM	**New Issues**	

MEETING OF NHGRI-SUPPORTED INVESTIGATORS ENGAGED IN LARGE-SCALE SEQUENCING

Hyatt Regency Bethesda
December 18-19, 1997

Participants

Mark Adams, Ph.D.
The Institute for Genomic Research

Maria Athanasiou, Ph.D.
University of Texas Southwestern
Medical School

Elbert Branscomb, Ph.D.
Lawrence Livermore National
Laboratory

Aravinda Chakravarti, Ph.D.
Case Western Reserve University

Sandra Clifton, Ph.D.
University of Oklahoma

Francis Collins, M.D., Ph.D.
NHGRI

David Cox, Ph.D.
Stanford University

Pieter de Jong, Ph.D.
Roswell Park Cancer Institute

Glen Evans, M.D., Ph.D.
University of Texas Southwestern
Medical School

Adam Felsenfeld, Ph.D.
NHGRI

Marvin Frazier, Ph.D.
Department of Energy

Richard Gibbs, Ph.D.
Baylor College of Medicine

Phil Green, Ph.D.
University of Washington

Mark Guyer, Ph.D.
NHGRI

Elke Jordan, Ph.D.
NHGRI

Francis Kalush, Ph.D.
The Institute for Genomic Research

Bruce Kimmel, Ph.D.
University of California, Berkeley

Eric Lander, Ph.D.
Whitehead Institute

Charles Langley, Ph.D.
University of California, Davis

Richard Mathies, Ph.D.
University of California, Berkeley

Richard Myers, Ph.D.
Stanford University

De Lill Nasser, Ph.D.
National Science Foundation

David Nelson, Ph.D.
Baylor College of Medicine

Maynard Olson, Ph.D.

University of Washington

Jane Peterson, Ph.D.
NHGRI

Rudy Pozzatti, Ph.D.
NHGRI

Bruce Roe, Ph.D.
University of Oklahoma

Gerald Rubin, Ph.D.
University of California, Berkeley

Jeffery Schloss, Ph.D.
NHGRI

Hiroaki Shizuya, Ph.D.
California Institute of Technology

Clark Tibbetts, Ph.D.
George Mason University

Bob Waterston, Ph.D.
Washington University

Bob Weiss, Ph.D.
University of Utah

Rick Wilson, Ph.D.
Washington University

Barbara Wold, Ph.D.
California Institute of Technology

METHODS FOR DISCOVERING AND SCORING
SINGLE NUCLEOTIDE POLYMORPHISMS

NIH GUIDE,

RFA: HG-98-001

P.T.

National Institutes of Health
NCI NCRR NEI NHGRI NHLBI NIA NIAAA NIAID NIAMS NICHD
NIDA NIDCD NIDDK NIDR NIEHS NIGMS NIMH NINDS

Letter of Intent Receipt Date: March 25, 1998
Application Receipt Date: May 7, 1998

PURPOSE

The purpose of this RFA is to solicit applications for research grants to (1) develop genomic-scale technologies, or (2) implement pilot-scale or large-scale projects for the discovery and scoring of single nucleotide polymorphisms (SNPs). The pilot/large-scale projects may be for SNPs that are located throughout the genome or that are located in particular genome regions or in sets of genes related to particular processes, organs, or diseases. The availability of a dense collection of SNPs will stimulate many areas of biological research, including the identification of the genetic components of disease.

HEALTHY PEOPLE 2000

The Public Health Service (PHS) is committed to achieving the health promotion and disease prevention objectives of "Healthy People 2000," a PHS-led national activity for setting priority areas. This Request for Applications (RFA), "Methods for Discovering and Scoring Single Nucleotide Polymorphisms", is related to several priority areas, including cancer, heart disease and stroke, diabetes and chronic disability conditions, maternal and infant health, and others. Potential applicants may obtain a copy of "Healthy People 2000" (Full Report: Stock No. 017-001-00474-0) or "Healthy People 2000" (Summary Report: Stock No. 017-001-00473-1) through the Superintendent of Documents, Government Printing Office, Washington, DC 20402-9325 (telephone 202-783-3238).

ELIGIBILITY REQUIREMENTS

Applications may be submitted by domestic and foreign for-profit and non-profit organizations, public and private organizations, such as universities, colleges, hospitals, laboratories, companies, units of State and local governments, and eligible agencies of the Federal Government. Applications from social/ethnic minority individuals, women, and persons with disabilities are encouraged.

MECHANISM OF SUPPORT

All of the institutes participating in this RFA will use the National Institutes of Health (NIH) individual research grant (R01). In addition, several of the institutes will use the program project grant (P01) or the pilot project/feasibility study (R21) mechanisms; investigators considering applying for either an R21 or P01 grant should contact the appropriate program officer (see below). The total project period for R01 and P01 applications submitted in response to the present RFA may not exceed 3 years. The direct cost per year for R01 or P01 grants may not exceed $500,000 without prior discussion with the relevant program officer.

Responsibility for the planning, direction and execution of the proposed project will be solely that of the applicant. Awards will be administered under PHS grants policy as stated in the Public Health Service Grants Policy Statement. Future unsolicited competing continuation applications will compete with all investigator-initiated applications and will be reviewed according to the customary peer review procedures.

All applications received in response to this solicitation will, for administrative reasons, be assigned initially to NHGRI. After discussions among the participating Institutes and Centers, applications will be reassigned to the Institute(s) or Center(s) that are programmatically most appropriate. Because the scope of the research proposed in response to this RFA encompasses the interests of several NIH Institutes and Centers, applications may receive dual assignments based on the established PHS referral guidelines. Awards will be made and managed by the NHGRI and/or the other participating Institutes and Centers. The earliest anticipated award date is September 30, 1998.

FUNDS AVAILABLE

It is anticipated that $10 million per year will be available for this initiative. Awards pursuant to this RFA are contingent upon the availability of funds for this purpose. The amount of funding for these projects may be increased if a large number of highly meritorious applications are received and if funds are available. Only applications that are found to be of high scientific merit will be considered for funding and not all of the funds will be spent if there are not enough highly meritorious applications. Funding in future years will be subject to the availability of funds.

RESEARCH OBJECTIVES

Background

Genetic factors appear to contribute to virtually every human disease, conferring susceptibility or resistance, affecting the severity or progression of disease, and interacting with environmental influences. Much of current biomedical research, in both the public and private sectors, is based upon the expectation that understanding the genetic contribution to disease will revolutionize diagnosis, treatment, and prevention. Defining and understanding the role played by genetic factors in disease will also allow the non-genetic, environmental influence(s) on disease to be more clearly identified and understood.

Analysis of DNA sequence variation is becoming an increasingly important source of information for identifying the genes involved in both disease and in normal biological processes, such as development, aging, and reproduction. In trying to understand disease processes, information about genetic variation is critical for understanding how genes function or malfunction, and for understanding how genetic and functional variation are related. Response to therapies can also be affected by genetic differences. Information about DNA sequence variation will thus have a wide range of application in the analysis of disease and in the development of diagnostic, therapeutic, and preventative strategies.

Completion of the first human DNA sequence, through the efforts of the Human Genome Project (HGP), is expected by 2005. While this will be of immense significance for many reasons, the HGP will actually produce very little information about DNA sequence variation within the human population. Although the DNA sequence that will be produced by the HGP will come from several individuals, at most positions the sequence will come from only one. The exceptions will be regions where overlapping clones from different chromosomes will be sequenced, but such overlap will be less than 10% of the complete sequence. Even in the overlap regions, DNA from only two chromosomes will be represented at any given site. Thus, additional studies are needed to discover the amount and distribution of variation in human DNA.

There are several types of DNA sequence variation, including insertions and deletions, differences in the copy number of repeated sequences, and single base pair differences. The latter are the most frequent. They are termed single nucleotide polymorphisms (SNPs) when the variant sequence type has a frequency of at least 1% in the population. SNPs have many properties that make them attractive to be the primary analytical reagent for the study of human sequence variation. In addition to their frequency, they are stable, having much lower mutation rates than do repeat sequences. Detection methods for SNPs are potentially more amenable to being automated and used for

large-scale genetic analysis. Most importantly, the nucleotide sequence variations that are responsible for the functional changes of interest will often be SNPs.

As noted, SNPs are very common in human DNA. Any two random chromosomes differ at about 1 in 1000 bases. For any particular polymorphic base (i.e., a base where the least common variant has a frequency of at least 1% in the population), only half or fewer of random pairs of chromosome differ at that site. Thus, there are actually more sites that are polymorphic in the human population, viewed in its entirety, than the number of sites that differ between any particular pair of chromosomes. Altogether, there may be anywhere from 6 million to 30 million nucleotide positions in the genome at which variation can occur in the human population. Thus, overall, approximately one in every 100 to 500 bases in human DNA may be polymorphic.

Information about SNPs will be used in three ways in genetic analysis. First, SNPs can be used as genetic markers in mapping studies. SNPs can be used for whole-genome scans in pedigree-based linkage analysis of families. A map of about 2000 SNPs has the same analytical power for this purpose as a map of 800 microsatellite markers, currently the most frequently used type of marker. Second, when the genetics of a disease are studied in individuals in a population, rather than in families, the haplotype distributions and linkage disequilibria can be used to map genes by association methods. For this purpose, it has been estimated that 30,000 to as many as 300,000 mapped SNPs will be needed.

Third, genetic analysis can be used in case-control studies to directly identify functional SNPs contributing to a particular phenotype. Because only 3-5% of the human DNA sequence encodes proteins, most SNPs are located outside of coding sequences. But SNPs within protein-coding sequences (which have recently been termed cSNPs) are of particular interest because they are more likely than a random SNP to have functional significance. It is also undoubtedly the case that some of the SNPs in non-coding DNA will also have functional consequences, such as those in sequences that regulate gene expression. Discovery of SNPs that affect biological function will become increasingly important over the next several years, and will be greatly facilitated by the availability of a large collection of SNPs, from which candidates for polymorphisms with functional significance can be identified. Accordingly, discovery of a large number of SNPs in human DNA is one objective of this RFA.

SNPs will be particularly important for mapping and discovering the genes associated with common diseases. Many processes and diseases are caused or influenced by complex interactions among multiple genes and environmental factors. These include processes involved in development and aging, and common diseases such as diabetes, cancer, cardiovascular and pulmonary disease, neurological diseases, autoimmune diseases, psychiatric illnesses, alcoholism, common birth defects, and susceptibility to infectious diseases, teratogens, and environmental agents. Many of the alleles associated with health problems are likely to have low penetrance, meaning that only a few of the individuals carrying them will develop disease. However, because such polymorphisms are likely to be very common in the population, they make a significant contribution to the health burden of the population. Examples of common polymorphisms associated with an increased risk of disease include the ApoE4 allele and Alzheimer's disease, and the APC I1307K allele and colon cancer.

Most of the successes to date in identifying (a) the genes associated with diseases inherited in a Mendelian fashion, and (b) the genetic contribution to common diseases, e.g. BRCA1 and 2 for breast cancer, MODY 1, 2, and 3 for type 2 diabetes, and HNPCC for colon cancer, have been of genes with relatively rare, highly penetrant variant alleles. These genes are well-suited to discovery by linkage analysis and positional cloning techniques. However, the experimental techniques and strategies useful for finding the low penetrance, high frequency alleles involved in disease are usually not the same, and are not as well developed, as those that have been successfully applied in positional cloning. For example, pedigree analysis of families often does not have sufficient power to identify common, weakly contributing loci. The types of association studies that do have the power to identify such loci efficiently require new approaches, techniques, and scientific resources to make them as robust and powerful as positional cloning. Among the resources needed is a genetic map of much higher density than the existing, microsatellite-based map. Association studies using a dense map should allow the identification of disease alleles even for complex diseases. SNPs are well suited to be the basis of such a map.

Available technologies have been used to discover SNPs with a reasonable degree of success. Thus, there is an opportunity to begin to test the feasibility of applying these methods in a high throughput, large-scale fashion to discover large numbers of SNPs. At the same time, there is clearly a need to improve these methods and to develop

new approaches to SNP discovery. Current methods for the discovery of SNPs are often not particularly appropriate to score known SNPs in genotyping assays, and the available scoring techniques leave much to be desired in terms of throughput, efficiency and cost. Thus, there is also a critical need to develop new methods for scoring known SNPs.

Technology development spans a spectrum of stages. Initially it involves the development of a new methodology or the significant improvement of an existing methodology to the point of proof of principle. The method must then be reduced to practice. For such a new method to have a significant impact on genomic studies, it must also be shown that it can be used efficiently on a large-scale or genomic basis; this requires another level of technology development. This RFA is intended to solicit applications that address any of these phases of technology development. Specifically, this RFA is intended to solicit research projects of two types: (1) development of new or improved methods, and (2) pilot-scale or large-scale projects, for SNP discovery and scoring. Of particular interest are technologies that can be applied at the "genomic scale" cost-efficiently, and can be easily exported into other laboratories, or in other ways made readily accessible to investigators.

Objectives and Scope

The tools needed to discover and score SNPs efficiently are just beginning to emerge and many more robust technologies are needed. The Human Genome Project has been successful in generating information and resources rapidly and economically, in part, by developing and applying high-throughput and efficient technologies. Therefore, the NIH seeks the development of technologies that can be applied in similar ways to the rapid and efficient discovery of SNPs and the scoring of SNPs in many samples. Large-scale projects for SNP discovery will allow comparison of the various existing technologies, particularly with respect to scalability, and will begin to generate a large collection of SNPs.

Applications are solicited in these areas:

1. Development of new or improved methods for high throughput, cost-efficient discovery or scoring (or both) of SNPs. SNP "discovery" involves finding new SNPs. SNP "scoring" involves methods to determine the genotypes of many individuals for particular SNPs that have already been discovered. Methods that involve "wet bench" approaches, computational approaches, or multiplexing are appropriate. Proposed methods may focus on obtaining SNPs throughout the genome, or may focus on cSNPs; they may also target particular types or sets of genes. Methods that yield additional information (e.g. map location, haplotypes) at the same time as the SNP itself are appropriate, although the costs and benefits of obtaining the additional information must be discussed. Applicants who propose to develop new methods for SNP discovery or scoring should discuss the potential advantages of the proposed methods over existing methods.

2. Pilot-scale or large-scale projects for SNP discovery, scoring, or both.

Pilot-scale or large-scale projects may be proposed that target random SNPs or cSNPs on a genome-wide basis, or all of the SNPs within a defined region of one to several megabases. Applications may focus on genes involved in particular processes or diseases of interest to particular Institutes, as listed below. Methods that focus on finding SNPs in coding sequences or regulatory regions, or on finding SNPs for functional variants of genes, are of particular interest. However, the methods must be capable of being applied on a large scale. Proposals should include a discussion of error rates, costs, and ease of scale up.

Most of the Institutes and Centers participating in this RFA have interests in genes that are related to particular processes, organs, or diseases, as listed below. In addition, some are interested in supporting development of methods that are either general or specific to genes in which they are interested, as noted below. Applications that propose to identify SNPs in or around genes of particular interest to a participating Institute are particularly welcome.

NCI - Genes involved in cancer.

NCRR - Genes and non-coding regions anywhere in the genome.

NEI - Genes involved in the development, function, and diseases of the eye.

NHGRI - Genes and non-coding regions anywhere in the genome.

NHLBI - Genes involved in the development, function, regulation, and diseases of the cardiovascular, pulmonary, and hematological systems.

NIA - Genes for repair enzymes for DNA, proteins, and lipids; antioxidant enzymes; apoptosis-related proteins; receptors; stress response proteins; transcription factors and neurodegenerative diseases of aging. Specific gene region: the WRN gene for Werner's syndrome.

NIAAA - Genes involved in function of the central nervous system, e.g., those encoding neurotransmitter receptors, transporters, and biosynthetic enzymes, neurotrophic factors and their receptors, ion channels, signal transducing proteins, and transcription factors. Genes whose products mediate the toxic effects of alcohol.

NIAID - Genes involved in susceptibility to infectious diseases, allergy, and autoimmunity.

NIAMS - Genes involved in arthritis and musculoskeletal and skin diseases.

NICHD - Genes involved in developmental biology, gametogenesis, fertilization, embryogenesis, organogenesis, and reproductive endocrinology; genes associated with the formation of birth defects; genes involved in mental retardation, autism and other developmental disabilities; genes associated with learning, behavior, and temperament; and genes affecting drug metabolizing enzymes in children.

NIDA - Genes involved in drug abuse and addiction.

NIDCD - Genes related to normal and disordered mechanisms of communication, including hearing, balance, voice, speech, language, taste and smell.

NIDDK - Genes involved in diabetes and digestive and kidney diseases.

NIDR - Genes involved in the development, function, and diseases of craniofacial, oral, and dental tissues.

NIEHS - Genes controlling the distribution and metabolism of toxicants; genes for DNA repair pathways; genes for the cell cycle control system; genes for cell death and differentiation; and genes for the signal transduction systems controlling expression of the genes in the other categories. For NIEHS, participation in this RFA is the first phase of the Environmental Genome Project; see http://dir.niehs.nih.gov/dirosd/policy/egp.html.

NIGMS - Genes and non-coding regions not targeted to disease.

NIMH - Genes involved in behavior, mental disorders, and the development, function, and regulation of the central nervous system.

NINDS - Genes involved in neurological processes, in particular those genes or chromosomal regions identified as related to neurological disorders or stroke.

The following Institutes and Centers are interested in supporting the development of methods that are either general or specific to genes in which they are interested - NCI, NCRR, NHGRI, NHLBI, NIA, NIAAA, NIAID, NIDA, NIDDK, NIDR, NIEHS, NIGMS, and NIMH.

Population Resources. Most genetic variation occurs within rather than between ethnic groups; this means that sequence variants that are common in one group are likely to be found in other groups as well. Efforts are currently under way to establish a central repository of anonymous DNA samples as a resource for the discovery of SNPs. This resource may be available by the time applications are funded under this RFA. However, applicants should propose one or more alternative sources of appropriate samples in case the planned resource is not available by that time. Applicants for SNP discovery projects should provide plans that will allow the detection of SNPs that are

common in the U.S. population. In most populations studied, the minimum frequency should be 1% for cSNPs and 10% for SNPs that are not in coding regions.

Human Subjects Issues Associated with SNP Discovery. Recently it has become evident that human subjects issues are raised by the large-scale sequencing of human genomic DNA because large amounts of DNA sequence information from single individuals will be generated. These issues are discussed in "Guidance on Human Subjects Issues in Large-Scale DNA Sequencing," which can be found on the NHGRI Home Page (http://www.NHGRI.nih.gov/Grant_info/Funding/Statements/large_scale.html). As a result of the research supported under this RFA, it is possible that an analogous situation might exist, i.e., that enough information might be developed about the genotypes of the individuals whose DNA was used to discover SNPs to allow them to be identified and, consequently, become subject to any risk(s) that might arise as a result of that identification. Applicants should address any special human subjects issues that arise as a result of their proposed research.

Data and Materials Dissemination. The sharing of materials, data, and software in a timely manner has been an essential element in the rapid progress that has been made in genome research. While Public Health Service (PHS) policy requires that investigators make unique research resources, including DNA sequences and mapping information, readily available when they have been published (PHS Grants Policy Statement, April 1, 1994, pp. 8-25 to 8-26), the advisors to the NIH and the Department of Energy (DOE) genome programs have encouraged more rapid sharing. This has, in fact, become the norm in the genome community.

NIH is interested in ensuring that the information about SNPs that is developed through this RFA becomes readily available to the research community for further research and development, in the expectation that this will eventually lead to products of benefit to the public. For this reason, NIH is concerned that patent applications on large numbers of SNPs, in the absence of such demonstrated utility, might have a chilling effect on the future development of products that can improve the public health. At the same time, NIH recognizes the rights of grantees to elect and retain title to subject inventions developed under Federal funding under the provisions of the Bayh-Dole Act. Indeed, for inventions developed in its intramural program, NIH does file patent applications, in accord with a set of policies that are described at http://www.nih.gov/od/ott/200po6.htm.

To address the joint interests of the government in the availability of, and access to, the results of publicly funded research and in the opportunity for economic development based on those results, NIH requires applicants who respond to this RFA to develop and propose specific plans for sharing the data, materials, and software generated through the grant. For this purpose, it is the opinion of the NIH that dissemination of such developments via individual laboratory web sites is not sufficient, as it would force interested investigators to have to search several different data collections to make use of the results of this initiative. It is preferable that data pertaining to all SNPs discovered or scored should be placed in a common, public database. Any additional information known, such as map location, should similarly be deposited in that database. A specific database suitable for this purpose will be identified when the awards are made.

The initial review group will comment on the proposed plan for sharing and data release. The adequacy of the plan will also be considered by NIH staff as one of the criteria for award. The proposed sharing plan, after negotiation with the applicant when necessary, will be made a condition of the award. Evaluation of renewal applications will include assessment of the effectiveness of data, material, and software release.

Applicants are also reminded that the grantee institution is required to disclose each subject invention to the Federal Agency providing research funds within two months after the inventor discloses it in writing to grantee institution personnel responsible for patent matters. The awarding Institute or Center reserves the right to monitor grantee activity in this area to ascertain if patents on large numbers of SNPs of ill-defined functionality are being filed.

Where appropriate, grantees may work with the private sector to make unique resources available to the larger biomedical research community at a reasonable cost. Applicants may request funds to defray the costs of sharing materials or submitting data, with adequate justification.

POST-AWARD MANAGEMENT

During the course of the grant period, it is anticipated that technologies will improve and the rate of progress and focus of work supported by the grant(s) may change. Accordingly, it is expected that the principal investigator(s) will make any necessary adjustments in scientific direction to accommodate such changes. During the course of the award period, the principal investigators may be invited to meet with NIH program staff in Bethesda, MD, to review scientific progress. Other scientists external to and knowledgeable about these studies may also be invited to participate. Budget requests should include travel funds for the P.I. to meet annually in the Washington D.C. area, should such meetings be advisable.

LETTER OF INTENT

Prospective applicants are strongly encouraged to discuss their research objectives and the appropriate grant mechanism with NIH staff in the relevant Institute or Center early in their planning process. Prospective applicants are asked to submit, by March 25, 1998, a letter of intent that includes a descriptive title of the proposed research, the name, address, e-mail address, and telephone number of the principal investigator, and the identities of other key personnel and participating institutions. Although a letter of intent is not required, is not binding, and does not enter into the review of subsequent applications, the information that it contains will allow NIH staff to estimate the potential review workload and to avoid conflict of interest in the review. The letter of intent should be sent to:

Lisa D. Brooks, Ph.D.
Program Director, Division of Extramural Research
National Human Genome Research Institute
Building 38A, Room 614
38 Library Drive MSC 6050
Bethesda, MD 20892-6050
[redacted]
Lisa_Brooks@nih.gov

INCLUSION OF WOMEN AND MINORITIES IN RESEARCH INVOLVING HUMAN SUBJECTS

It is the policy of the NIH that women and members of minority groups and their subpopulations must be included in all NIH supported biomedical and behavioral research projects involving human subjects, unless a clear and compelling rationale and justification is provided that inclusion is inappropriate with respect to the health of the subjects or the purpose of the research. This new policy results from the NIH Revitalization Act of 1993 (Section 492B of Public Law 104-43) and supersedes and strengthens the previous policies (Concerning the Inclusion of Women in Study Populations, and Concerning the Inclusion of Minorities in Study Populations) which have been in effect since 1990. The new policy contains some new provisions that are substantially different from the 1990 policies.

All investigators proposing research involving human subjects should read the "NIH Guidelines for Inclusion of Women and Minorities as Subjects in Clinical Research," which have been published in the Federal Register of March 28, 1994 (FR 59 14508-14513), and reprinted in the NIH GUIDE FOR GRANTS AND CONTRACTS of March 18, 1994, Volume 23, Number 11.

Investigators may obtain copies from these sources or from program staff or contact person listed under INQUIRIES. Program staff may also provide additional relevant information concerning the policy.

APPLICATION PROCEDURES

The research grant application form PHS 398 (rev. 5/95) is to be used in applying for these grants. These forms are available at most institutional offices of sponsored research or from the Office of Grants Inquiries, Room No. 1040, 6701 Rockledge Drive, Center for Scientific Review, National Institutes of Health, Bethesda, MD 20892, telephone ([redacted] and from the program officer listed below.

The RFA label available in the PHS 398 (rev. 5/95) application form must be affixed to the bottom of the face page of the application. Failure to use this label could result in delayed processing of the application such that it may not reach the review committee in time for review. In addition, the RFA title and number must be typed on line 2a of the face page of the application form and the YES box must be marked.

Submit a signed, typewritten original of the application and three signed photocopies, in one package to:

Center for Scientific Review
National Institutes of Health
Room 1040
6701 Rockledge Drive
Bethesda, MD 20892
(Express Mail zip code is 20817)

At the time of submission, two additional copies of the application, including appendices, must also be sent to:

Dr. Rudy Pozzatti
Office of Scientific Review
National Human Genome Research Institute
Building 38A, Room 613
38 Library Drive, MSC 6050
Bethesda, MD 20892-6050

Applications must be received by May 7, 1998. If an application is received after that date, it will be returned to the applicant without review. The Center for Scientific Review (CSR) will not accept any application in response to this announcement that is essentially the same as one currently pending initial review, unless the applicant withdraws the pending application. The CSR will also not accept any application that is essentially the same as one already reviewed. This does not preclude the submission of substantial revisions of applications already reviewed, but such applications must include an introduction addressing the previous critique. The applicants should also ensure that their revised applications respond to the review criteria by which the applications in response to this RFA will be evaluated.

REVIEW CONSIDERATIONS

Upon receipt, applications will be reviewed for completeness by CSR and for responsiveness to the RFA by NIH program staff. Incomplete applications will be returned to the applicant without further consideration. If the application is not responsive to the RFA, NIH staff will contact the applicant to determine whether to return the application to the applicant or submit it for review in competition with unsolicited applications at the next review cycle.

Those applications that are complete and responsive will be evaluated for scientific and technical merit in accordance with the criteria stated below by an appropriate peer review group convened by the NHGRI. As part of the initial merit review, all applications will receive a written critique and may undergo a process in which only those applications deemed to have the highest scientific merit will be discussed and assigned a priority score. All applications will receive a second level of review by the appropriate National Advisory Council.

Review criteria will include:

o Significance: For technology development proposals, does this application address the development of a promising technology that can be usefully applied to the rapid and efficient discovery or scoring of SNPs? If the aims of the application are achieved, how will it improve the capabilities of researchers to discover SNPs or use SNPs in the genetic analysis of complex traits?

o For pilot-scale/large scale SNP discovery proposals, does this application address the efficient and rapid development of a useful resource of SNPs? If the aims of the application are achieved, how much will the SNP

collection that is available to the research community be improved?

o Approach: Are the conceptual framework, design, methods, and analyses appropriate and adequate to accomplish the aims of the project? For pilot-scale/large-scale projects, are the methods adequate to allow the rapid, efficient detection of SNPs? Does the applicant acknowledge potential problem areas and consider alternate approaches? Is the scientific and technical merit of the proposed research sufficient to advance the objectives of the RFA?

o Innovation: Does the project employ novel concepts, approaches or method? Are the aims original and innovative? Does the project propose to develop new or significantly improved methodologies or technologies for SNP discovery or scoring?

o Investigator: Are the Principal Investigator and staff appropriately trained and well suited to carry out this work? Is the work proposed appropriate to the experience level of the principal investigator and other researchers (if any)?

o Scalability: For technology development or pilot-scale SNP production projects, what is the likelihood that the technology or approach will be able to be used efficiently at a full production level in a timely manner?

o Environment: Does the scientific environment in which the work will be done contribute to the probability of success? Do the proposed experiments take advantage of unique features of the scientific environment or employ useful collaborative arrangements? Is there evidence of institutional support?

o Budget and duration: Are the proposed budget and duration appropriate in relation to the proposed research?

The availability of special opportunities for furthering research programs through the use of unusual talent resources, populations, or environmental conditions in other countries which are not readily available in the United States or which provide augmentation of existing U.S. resources will be considered in the review.

The initial review group also will examine the provisions for the protection of human and animal subjects, and the safety of the research environment.

For R21 applications, preliminary data are not required. However, the applicant does have the responsibility to develop a sound research plan and to present any other information that can be considered as evidence of feasibility.

The initial review group will also be asked to comment on the plans for making the data and materials developed under the proposed project accessible to the biomedical research community: Will the forthcoming methodologies, resources, software, and collections of SNPs be usable by, and accessible to, the broad scientific community of biomedical researchers who are discovering and using SNPs in a wide range of research investigations? Any opinions expressed by the reviewers about this aspect of the proposal will be recorded as an administrative note.

AWARD CRITERIA

The earliest anticipated date of award is September 30, 1998. Subject to the availability of funds, and consonant with the priorities of this RFA, the participating Institutes and Centers will provide funds for a project period of up to three years. Factors that will be used to make award decisions are:

* quality of the proposed project as determined by peer review;

* balance among the projects in addressing different experimental approaches and their complementarity to other ongoing efforts;

* adequacy of plans to make data and material developed as a result of the proposed research accessible to the biomedical research community in a timely manner; and

* availability of funds.

INQUIRIES

Written, telephone, and e-mail inquiries concerning this RFA are encouraged. The opportunity to clarify any issues or questions from potential applicants is welcome. Direct inquiries regarding programmatic issues and mechanisms of support to the following NIH staff.

NHGRI Lisa D. Brooks, Ph.D.
Program Director, Division of Extramural Research
National Human Genome Research Institute
Building 38A, Room 614
38 Library Drive MSC 6050
Bethesda, MD 20892-6050
[redacted]

NCI Grace L. Shen, Ph.D.
Program Director, Tumor Genetics Program
Cancer Genetics Branch
Division of Cancer Biology
National Cancer Institute
Executive Plaza North, Room 501
6130 Executive Boulevard
Rockville, MD 20892-7381
[redacted]

NCRR Marjorie Tingle, Ph.D
Health Science Administrator
National Center for Research Resources
Rockledge Center 1, Room 6154
6705 Rockledge Drive
Bethesda, Maryland 20892
[redacted]

NEI Maria Y. Giovanni, Ph.D
National Eye Institute, NIH
EPS Suite 350 MSC 7164
Bethesda, Maryland 20892-7164
[redacted]

NHLBI Stephen C. Mockrin, Ph.D.
Deputy Director, Division of Heart and Vascular Diseases
National Heart, Lung, and Blood Institute
National Institutes of Health
TWO Rockledge Centre, STE 9044
6701 Rockledge Drive MSC 7940
[redacted]

NIA Huber R. Warner, PhD
Deputy Associate Director
Biology of Aging Program
National Institute on Aging
Gateway Bldg, Room 2C231
Bethesda, MD, 20892
[redacted]

NIAAA Robert W.Karp, Ph.D.
Division of Basic Research
6000 Executive Boulevard, Suite 402,
MSC 7003
Bethesda, MD 20892-7003
[redacted]

NIAID Vicki Seyfert, Ph.D.
Chief, Immunoregulation Branch
Division of Allergy, Immunology and Transplantation
National Institute of Allergy and Infectious Diseases
Solar Building, 4A21
6003 Executive Blvd.
Bethesda, MD 20852
[redacted]
[redacted]@[redacted].gov

NIAMS Steven J. Hausman, Ph.D.
Director, Extramural Program
National Institute of Arthritis and Musculoskeletal and Skin Diseases
45 Center Drive MSC 6500
Building 45, Room 5AS-13F
Bethesda, Maryland 20892-6500
[redacted]
[redacted](fax)
[redacted]v

NICHD A. Tyl Hewitt, Ph.D.
Chief, Developmental Biology,
Genetics and Teratology Branch
National Institute of Child Health and Human Development
6100 Bldg., Room 4B01
9000 Rockville Pike
Bethesda, MD 20892
[redacted]

NIDA Theresa Lee, Ph.D.
Division of Basic Research
National Institute on Drug Abuse

5600 Fishers Lane, Room 10A19
Rockville, MD 20857
[redacted]

NIDDK Catherine McKeon, Ph.D.
Director, Metabolic Diseases and Gene Therapy Research Program
National Institute of Diabetes and Digestive and Kidney Diseases
Bldg 45 Rm 5AN.18B
Bethesda, MD 20892-6600
[redacted]

NIDCD Rochelle K. Small, Ph.D.
Health Science Administrator, Division of Human Communication
National Institute on Deafness and Other Communication Disorders
National Institutes of Health
EPS Room 400C
6120 Executive Blvd, MSC 7180
Bethesda, MD 20892-7180
[redacted]

NIDR Linda Thomas, Ph.D.
Director, Inherited Diseases and Disorders
Division of Extramural Research
National Institute of Dental Research
National Institutes of Health
Building 45 Room 4AN-24J
Bethesda, MD 20892-6402
[redacted]

NIEHS Jose M. Velazquez, Ph.D.
Program Administrator
National Institute of Environmental Health Sciences
National Institutes of Health
POB 12233
Research Triangle Park, NC
[redacted]

NIGMS Irene Eckstrand, Ph.D.
Program Director
Developmental and Cellular Processes Branch
National Institute of General Medical Sciences
45 Center Drive, Room 2AS.25K
Bethesda, MD 20892-6200
[redacted]

NIMH Steven O. Moldin, Ph.D.
Division of Basic and Clinical Neuroscience Research
National Institute of Mental Health
5600 Fishers Lane, Room 10C-26
Rockville, MD 20857
[redacted]

NINDS Judy Small, Ph.D.
Health Scientist Administrator
Division of Fundamental Neurosciences and Developmental Disorders
National Institute of Neurological Disorders and Stroke
Federal Building, Room 8C04
Bethesda, MD 20892
[redacted]

Direct inquiries regarding fiscal matters to:

Ms. Jean Cahill
Grants Management Officer
National Human Genome Research Institute
Building 38A, Room 613
38 Library Drive, MSC 6050
Bethesda, MD 20892-6050
[redacted]

AUTHORITY AND REGULATIONS

This program is described in the Catalog of Federal Domestic Assistance No. 93.172. Awards are made under authorization of the Public Health Service Act, Title IV, Part A (Public Law 78-410, as amended by Public Law 99-158, 42 USC 241 and 285) and administered under PHS grants policies and Federal Regulations 42 CFR 52 and 45 CFR Part 74. This program is not subject to the intergovernmental review requirements of Executive Order 12372 or Health Systems Agency review.

The Public Health Service (PHS) strongly encourages all grant recipients to provide a smoke-free workplace and promote the non-use of all tobacco products. This is consistent with the PHS mission to protect and advance the physical and mental health of the American people.

Full Text HG-98-002

RESEARCH NETWORK FOR LARGE-SCALE SEQUENCING OF THE HUMAN GENOME

NIH GUIDE, JANUARY 9, 1998

RFA HG-98-002

P.T.

Keywords:

National Human Genome Research Institute

Letter of Intent Receipt Date: July 1, 1998
Application Receipt Date: October 9, 1998

PURPOSE

The purpose of this RFA is to seek applications to participate in a Research Network, the goal of which is to make a major contribution to the completion of the first human genome sequence by 2005. This Research Network will be comprised of sequence production centers, specialized sequencing projects and a quality control center.

HEALTHY PEOPLE 2000

The Public Health Service (PHS) is committed to achieving the health promotion and disease prevention objectives of "Healthy People 2000," a PHS-led national activity for setting priority areas. This RFA, Research Network for Large-Scale Sequencing of the Human Genome, is related to several priority areas including cancer, heart disease and stroke, diabetes and chronic disability conditions, and maternal and infant health. Potential applicants may obtain a copy of "Healthy People 2000" (Full Report: Stock No. 017-001-00474-0) or "Healthy People 2000" (Summary Report: Stock No. 017-001-00473-1) through the Superintendent of Documents, Government Printing Office, Washington, DC 20402-9325 (telephone 202-783-3238).

ELIGIBILITY REQUIREMENTS

Applications may be submitted by domestic non-profit and for-profit organizations, private and public, such as universities, colleges, companies, hospitals, laboratories, units of state or local governments, and eligible agencies of the Federal government. Applications from minority individuals and women are encouraged. Applications from foreign institutions will not be

accepted; however subcontracts to foreign institutions will be considered. A principal investigator submitting an application for the sequencing production centers must have demonstrated experience in directing projects that have produced at least 7.5 Mb of high quality finished DNA sequence.

MECHANISM OF SUPPORT

The administrative and funding mechanism to be used to support this program will be the Cooperative Agreement (U01), an "assistance" mechanism, which is distinguished from a regular research grant in that substantial scientific and/or programmatic involvement by NHGRI staff with the awardee is anticipated. The cooperative agreement is used when participation by NIH staff is warranted to support and/or stimulate the recipient's activity by involvement in and otherwise working jointly with the award recipient in a partner role; NIH staff will not assume direction, prime responsibility, or a dominant role in the activity. Details of the responsibilities, relationships, and governance of the studies funded under cooperative agreement(s) are discussed later in this document under the section "Terms and Conditions of Award." Each component of the Research Network will be awarded as a separate U01.

The Research Network will be composed of three separate, but complementary, activities: 1) sequence production centers; 2) specialized sequencing projects, and 3) a quality control center. The objectives of the three types of projects are described below. The project period that may be requested for each type of project is as follows: 1) up to five years for sequence production centers, 2) up to three years for specialized sequencing projects and 3) up to three years for the quality control center. Similarly, the sizes of the different types of awards will vary. The earliest anticipated award date is July 1, 1999. It is the intention of NHGRI that the Research Network will continue through Fiscal Year 2005, if needed to complete the human DNA sequence. NHGRI expects to solicit additional specialized sequencing projects during the term of the Research Network if funds are available and if continued activity of this type is warranted. NHGRI is committed to ongoing assessment of the quality of the DNA sequence produced in this project and therefore it is anticipated that there will be a future solicitation to continue a quality control center beyond the three-year term of the center that will be funded under this RFA.

FUNDS AVAILABLE

The estimated funds available for the first year of support for awards under this RFA will be $60 million per year (total costs) for three to five sequence production projects and at least $10 million per year (total costs) for up to four specialized sequencing projects and one quality control center.

The usual PHS policies governing grants administration and management will apply. This level of support is dependent on the receipt of a sufficient number of applications of high scientific merit. Beyond the first year, the funding level of each of the centers will be based on an annual evaluation. For the sequencing production centers, the evaluation criteria will be whether progress toward completion of the sequence of the human genome is sufficient and is state-of-

the-art, relative to that of the other sequence production centers, as determined by the Advisory Committee (see below), NHGRI staff and the National Advisory Council for Human Genome Research (NACHGR). For the specialized sequencing projects and the quality control center the criteria will be whether the project is meeting its goals and fulfilling the long- and short-term needs of the Research Network, as determined by the Advisory Committee, NHGRI staff and the NACHGR. The funding level for the Research Network will also be dependent upon the availability of funds.

RESEARCH OBJECTIVES

Background

The NHGRI is currently engaged, along with several other federal, private, and international organizations, in a fifteen-year research program called the Human Genome Project (HGP). The goals of the HGP are to characterize the genomes of human and selected model organisms through complete mapping and sequencing, to develop technologies for genomic analysis, to examine the ethical, legal, and social implications of human genetics research, and to train scientists who will be able to utilize the tools and resources developed through the HGP to pursue biological studies that will improve human health.

The HGP started in 1990 and significant progress toward completing these goals has been made in the past seven years; several goals have already been achieved. The genetic mapping goals for both the human and the mouse have been met. The human and mouse physical mapping goals are nearly complete. There has also been good progress toward meeting the sequencing goals. The DNA sequence of both the E. coli and S. cerevisiae genomes has been determined (as have those of several other microorganisms), the sequence of the C. elegans genome is expected to be finished by 1998, and the complete DNA sequence of D. melanogaster is expected to be finished early in the next century.

Producing the first reference human DNA sequence by 2005 is now the HGP's primary goal. In the early years of the HGP, the focus of the research program was on mapping and technology development because it was recognized that good maps and better technology were needed if the entire human DNA sequence was to be completed within the projected budget and time period. Three years ago, it was concluded that map construction and technology development had progressed sufficiently to warrant initiation of a pilot-scale sequencing program to develop and test approaches to full-scale production sequencing of human DNA. NHGRI funded six pilot projects for this purpose in 1996. Under the pilot project program, several different strategies have been implemented, a number of new technologies have been developed or implemented, and new informatics tools have been implemented to handle the data. In the course of developing their sequence capabilities, the pilot projects have deposited more than 30 Mb of high-quality mammalian genomic DNA sequence in GenBank.

The HGP goals also call for side-by-side sequencing of regions of the mouse genome syntenic with regions of the human genome because these sequences will help to inform the discussion of

the use of the mouse sequence in understanding the human sequence. The NHGRI-funded pilot projects have produced a few megabases of sequence from syntenic regions of the mouse genome.
In planning a program that will complete the sequence of the human genome by 2005, the most important component is adequate sequence production capacity. Additional infrastructural/organizational issues that need to be addressed in scaling up the current sequencing program effort have been identified during the pilot project program. It will be critical to ensure that the sequence production groups remain efficient and continue to evolve and become more efficient throughout the term of the program, and a number of additional supportive activities will be required to sustain and increase the productivity of the sequence production efforts over the long term. These include separate research efforts to: 1) evaluate new technology at a production level, 2) address problem regions in DNA (e.g., gaps, closure), 3) provide opportunities for new groups with promising approaches to attain production levels, and 4) evaluate the quality of the DNA sequence produced. It will also be important to facilitate continued and expanded communication and frequent exchange of information among the individual projects and NHGRI staff, as this has been found to provide significant benefits to the overall sequencing effort during the pilot project period.

Through this RFA, the NHGRI proposes to follow up the pilot project program with a Research Network that will make a substantial contribution to the international effort to sequence the human genome. Specifically, the Research Network's goal will be to complete 1.8 billion base pairs (60%) of human DNA sequence by 2005 (it is anticipated that the U.S. Department of Energy (DOE) and international partners in the HGP will complete the remaining 1.2 billion base pairs (40%) of the human genome sequence). This will require that the NHGRI program produces an annual average of almost 300 Mb of finished sequence, between 1999 and 2005. The coordination of the NHGRI effort by the Research Network is intended to enhance the productivity of the group as a whole, and thus increase the likelihood that the human sequence will be completed on time and within budget.

Completing the human sequence by 2005 will require the commitment of a substantial portion of NHGRI's resources. It is important to note, however, that NHGRI will continue to support the development of novel genomic technologies, including sequencing technology, outside of the Research Network, through its traditional grant program.

Research Objectives and Scope

As stated above, the goal of the Research Network will be to complete 1.8 billion base pairs (60%) of the first human DNA sequence by 2005. This RFA calls for three types of components to make up the Research Network that will accomplish this goal:

A. Sequence production centers: These projects will be the central DNA sequence production units of the network. Acceptable objectives for applications include production of at least 20 Mb of finished human sequence in the first year, at least 40 to 50 Mb in the second year and at least 50 to 100 Mb in each year thereafter, at a cost of no more than $0.40 (total costs) per base pair in

the first year and at an average of $0.25 or less per base pair over the life of the Research Network. Applicants for sequence production centers must follow the guidance given in the section below, entitled "Application Guidance for Production Sequencing." Proposals must also address the NHGRI policies for large-scale sequencing outlined below.

B. Specialized sequencing projects: The primary objective of this component of the Research Network is to increase the likelihood that the human DNA sequence will be completed by providing flexibility, capabilities or services that the sequence production centers cannot. Augmenting and complementing the sequence production centers, specialized sequencing projects can contribute to the overall HGP sequencing effort in any of a variety of ways. The following are examples of activities that would be appropriate for specialized sequencing projects:

- o testing one or more new sequencing technologies or strategies that have the potential, when implemented at large scale before 2005, to surpass the performance of those currently being used for large-scale production sequencing. These projects could be undertaken with the intent of exporting the technology or strategy, once it has been demonstrated to be robust, to a production center or of scaling it up to production levels at the test site with the intent of becoming a sequence production center;
- o serving as a service center to, for example, sequence difficult regions or close gaps.

This list is not intended to be inclusive and other ideas for specialized sequencing projects are welcome. All applications for specialized sequencing projects must present a plan, including a time line, that describes how and when the proposed effort will make a substantial contribution to the completion of the human DNA sequence. If a specialized project proposes to include a moderate-sized sequencing capacity in order to carry out its purpose, evidence of past experience in sequencing should be provided using the format available at (http://www.nhgri.nih.gov/DER/Announcements/progress_reports.html) The unit cost of sequencing in such a project should not exceed twice the average unit cost of sequencing in the production sequencing projects in the previous year. Applicants proposing to sequence at a moderate scale must also address issues listed below under "Application Guidance for Sequence Production Centers."

C. Quality control center: Applications are sought to support one cooperative agreement to evaluate, on an ongoing basis, the quality of the DNA sequence being produced by the sequence production and specialized sequencing centers. This is a new activity in the large-scale sequencing program. During the pilot project period, two sequence quality assessment exercises were completed; a description of the methods used is available (http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/). While the methods proposed for the quality control center need not be the same as those used in the previous exercises, applications that propose to carry out quality assessment must provide evidence for the robustness of the method(s) proposed or plans for assessing the validity of the proposed method(s) and of the sampling methods (e.g., how much material will be sampled). The first assessment of the quality of finished sequence must be completed no later than six months after funding of the quality control center, and the applicant should propose a plan for continued semi-

annual assessments of data from the production centers and specialized sequencing centers, where needed. It is expected that sequence quality assessment methods will evolve over the period of the grant; therefore the applicant should provide a plan to ensure that the methods being used will be maintained at the state-of-the-art. The quality control center must also include an outreach capability to provide advice to, and assist, the sequence production centers with their in-house quality control programs. The quality control center should also include funds to cover the cost of the Steering and Advisory Committees' activities.

A principal investigator may apply for more than one of the types of centers described above. However, no P.I. will be awarded a production center and a quality control center and it is unlikely that any other combination of two awards will be made to one P.I. although two awards may be made to one institution.

NHGRI POLICIES CONCERNING LARGE-SCALE SEQUENCING

During the past two years, as the pilot projects began to produce significant quantities of human DNA sequence, a number of issues arose that required the development of new policies by NHGRI. These policies will apply during the term of the Research Network. Thus, where appropriate, applicants must present plans to adhere to the policies.

Intellectual property. In NHGRI's opinion, in the absence of additional biological information, human genomic DNA sequence information should be freely available for use by the entire research community and, therefore, should not be patented but released into the public domain. NHGRI will monitor its grantees' activities with respect to patenting human genomic sequence. (see web site:http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/).

Data Release.
Finished mapping sequence and data: The U.S. HGP has adoped a policy of encouraging rapid release of mapping and sequence data into public databases. Guidelines developed by NHGRI and DOE advisors recommend that data be made publicly available within six months of the time they are verified (see http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/).

Unfinished sequence data: Participants in the international human DNA sequencing effort have recommended that early stage human sequence data should be rapidly released. In response, NHGRI determined that its grantees should release all sequence assemblies of 2,000 base pair units or larger within 24 hours of assembly (see http://www.nhgri.gov:80/Grant_info/Funding/Statements). Applicants should fully describe their plans for the release of mapping data and finished and unfinished sequence data.

Human subjects protection. Donors whose DNA will be sequenced in the project must give appropriate informed consent, and their confidentiality and anonymity must be ensured to the extent possible (recognizing that, because each individual's DNA sequence is unique, anonymity cannot ultimately be guaranteed). These issues have been addressed in a Guidance for the Use of DNA in Large-Scale Sequencing that was jointly issued by the NHGRI and the DOE human

genome program in August, 1996 (see http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/). It is expected sufficient libraries will be available by the anticipated award date of grants funded under this RFA (July 1, 1999); if these libraries are available, human genomic DNA sequence generated under this RFA MUST be determined from resources made according to the NHGRI-DOE Guidance.

Sequence quality. Quality standards are an important component of this program. After considerable discussion, the NHGRI adopted the goals that the sequence should be 99.99% accurate and there should be no gaps, either within or between clones (see http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/). Two sequence quality assessment exercises have been completed and have demonstrated that (1) it is possible to measure sequence quality at a low cost, and (2) that it is possible to produce sequence that meets the standard for accuracy. It is recognized, however, that it may not be possible to sequence all regions of the human genome to this standard. The policy recognizes this by providing that, in such regions, the sequence must be annotated to indicate what efforts were actually made to obtain high quality data. In the case of gaps, the annotation must include the size of the gap and the orientation of sequence fragments.

APPLICATION GUIDANCE FOR PRODUCTION SEQUENCING

Applicants must consider and address the following in preparing applications for sequence production projects called for in this RFA:

Progress Report. In order to achieve the ambitious first-year goal outlined above, applicants must already have a proven record for high-throughput DNA sequencing (At least 7.5 Mb of finished sequence. As the standards for sequence quality are evolving rapidly, for purposes of this RFA, the applicable standard of quality will be posted on the NHGRI Web site (http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/) as of 7/1/98; see Sequence Quality section below.). NHGRI has developed a progress report submission format that will allow applicants to submit large amounts of mapping and sequencing information electronically and that can be easily examined by the NHGRI staff and reviewers; this progress report form is available at http://www.nhgri.nih.gov/DER/Announcements/progress_reports.html . Evidence of the applicant's past sequencing accomplishments must be provided electronically using this form. The remaining components of the NIH application are not to be submitted through this electronic format and should be sent to NIH in the printed form called for in the NIH application kit. A printed copy of the textual material contained in the electronic progress report (excluding Part B, the graphical and tabular material) should also be included with the application.

Sequence Production Plan. The applicant must present a plan and propose milestones for achieving the proposed level of scale up. This plan must cover all phases of sequence production, starting with construction of a sequence-ready map, through deposition of the finished sequence in GenBank. Issues that should be discussed include: (1) the choice of regions to be sequenced, including any special considerations that may arise specifically because of that

choice, (2) the construction of sequence-ready maps, (3) sample preparation, (4) the sequencing process, (5) assembly of the finished sequence from the raw sequence traces and (6) automated annotation. It will be important to discuss bottlenecks or other problems that may be anticipated as the project increases in scale and how they will be addressed.

Sequence Cost. The calculated cost of sequencing (both prior and projected sequencing costs) must take into account all of the expenses associated with sequence production, beginning with construction of a sequence-ready map, through deposition of the finished sequence in GenBank (the costs of the sequence-ready maps must be included whether or not the maps are being produced in-house or at a different site). The total cost of sequencing must also include any production-related technology development (see below) that has been or will be supported by the project. However, the applicant may also provide a break-down of costs so that the reviewers can evaluate the contribution of different cost elements, such as production-related technology development, to the reported total cost.

Sequence Quality: Applicants must agree to submit their data for quality assessment during both the pre-award period (in order to allow the peer reviewers to evaluate this important factor) and during the course of the project period. For sequence data already finished by the application submission date, the assessment will be conducted using the methods employed in the previous NHGRI quality assessment exercises (http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/), in which all NHGRI grantees funded for large-scale sequencing cooperated in assessment of each other's data. This evaluation will be conducted after the application is submitted, but prior to the review meeting. The finished sequence to be evaluated will be chosen by NHGRI staff from the list of finished clones submitted by the P.I. in the progress report. During the course of the project period, the assessment will be done by the quality control center.

Each sequence production center must also implement an internal quality control program. Applicants must propose an internal quality control program that evaluates sequence accuracy, fidelity to the genome and short-range and long-range contiguity. If a program is already in use in the applicant's project, evidence of its usefulness must be presented in the Progress Report.

Production-related technology development/implementation. One of the goals of the NHGRI sequencing program is continually to improve the efficiency and decrease the cost of production sequencing. This will facilitate completion of the human genome sequence in the shortest possible time and at the lowest possible cost, as well as build the infrastructure that will be needed to continue sequencing multi-megabase regions of DNA from both the human and other organisms after the first human DNA sequence is completed. Applicants must present a plan to address this issue and discuss how their proposed project will balance further technology development and sequence production. As much as 10% of the requested budget may be budgeted for technology development/implementation for this purpose, but the cost must be included in the total cost of the sequence produced, as discussed above.

Mouse Genomic Sequence. While the focus of activity in the sequence production centers must

be on human genomic DNA sequence, as much as 10% of the effort may be devoted to sequencing genomic regions of the mouse that are syntenic with regions of the human genome that have already been or are being sequenced. The applicant must present a plan for obtaining the sequence-ready maps for the regions of the mouse to be sequenced and a coherent scientific strategy/rationale as to why the target regions were chosen.

Management Plan. The management of a sequencing center requires a significant commitment by the P.I. of the project. Accordingly, he or she is expected to devote at least 30% effort to the project. The applicant must propose a management plan for the project that takes into account the changes that will occur as the project scales up.

SPECIAL REQUIREMENTS FOR COOPERATIVE AGREEMENTS

I. Definitions

ARBITRATION PANEL: A panel that is formed to review scientific or programmatic disagreement (within the scope of the award) that may arise between award recipients and NHGRI. It will be composed of three members: a designee of the Steering Committee chosen without the NHGRI staff voting, one NHGRI designee, and a third designee with expertise in the relevant area who is chosen by the other two; in the case of an individual disagreement, the first member may be chosen by the individual awardee. The Arbitration Panel will help resolve both scientific and programmatic issues that develop during the course of work that restrict progress.

AWARDEE: The institution to which the cooperative agreement is awarded.

COOPERATIVE AGREEMENT: An assistance mechanism in which there is anticipated substantial NHGRI programmatic involvement with the recipient organization during the performance of the planned activity.

RESEARCH NETWORK: A group of scientists, each funded by a separate cooperative agreement, working together to complete the DNA sequence of the human genome by 2005.

NHGRI PROGRAM DIRECTOR(S): A scientist(s) of the NHGRI extramural staff who provides normal stewardship for the award and who, in addition, has substantial scientific/programming involvement during conduct of this activity, as defined in the terms and conditions of award. This involvement includes coordinating NHGRI's participation in the Research Network, functioning as a peer with the Principal Investigators, facilitating the partnership relationship between NHGRI and the Research Network, helping to maintain the overall scientific balance in the program commensurate with new research and emerging research opportunities, and ensuring that the Research Network program is consistent with the NHGRI missions and goals.

PRINCIPAL INVESTIGATOR (P.I.): The person who assembles the project, is responsible for

submitting the application in response to this RFA, and is responsible for the performance of the project. The Principal Investigator will coordinate project activities scientifically and administratively.

STEERING COMMITTEE (SC): A committee that is the main governing board of the Research Network. Membership includes the NHGRI Program Director(s), the P.I. of each awarded cooperative agreement (including the sequence production centers, specialized sequencing centers and the quality control center), and three research scientists with relevant expertise, but who are not affiliated with any of the projects participating in the Research Network.
SCIENTIFIC ADVISORY PANEL (AC): A committee that evaluates the progress of the Research Network and provides recommendations to the Director, NHGRI about continued support of the components of the Research Network. The Advisory Committee will be composed of four to six senior scientists with relevant expertise and who are not P.I.s of a cooperative agreement involved in the Research Network. The AC will meet at least annually.

II. Terms and Conditions of Award

The following terms and conditions will be incorporated into the award statement and will be provided to the Principal Investigator, as well as the appropriate institutional official, at the time of award. The following special terms of award are in addition to, and not in lieu of, otherwise applicable OMB administrative guidelines, HHS grant administration regulations at 45 CFR Parts 74 and 92 [Part 92 is applicable when State and local Governments are eligible to apply], and other HHS, PHS, and NIH grant administration policies:

1. The administrative and funding instrument used for this program will be the Cooperative Agreement (U01), an "assistance" mechanism (rather than an "acquisition" mechanism), in which substantial NIH scientific and/or programmatic involvement with the awardee is anticipated during the performance of the activity. Under the Cooperative Agreement, the NIH purpose is to support and/or stimulate the recipient's activity by involvement in and otherwise working jointly with the award recipient in a partner role, but it is not to assume direction, prime responsibility, or a dominant role in the activity. Consistent with this concept, the dominant role and prime responsibility for the activity resides with the awardee(s) for the project as a whole, although specific tasks and activities in carrying out the study will be shared among the awardee(s) and the NHGRI Program Director(s).

2. P.I. Rights and Responsibilities

The P.I. will have the primary responsibility for defining the details for the project within the guidelines of the RFA and for performing the scientific activity. The P.I. will agree to accept close coordination, cooperation, and participation of NHGRI staff in those aspects of scientific and technical management of the project as described under "NHGRI Program Staff Responsibilities".

The P.I. of a sequence production center will:

- Determine experimental approaches, design protocols, set project milestones and conduct experiments
- Produce genomic sequence to meet a quality standard and cost agreed upon at the time of award
- Release data according to NHGRI policies and publish results
- Submit data for quality assessment by the quality control center or in any other manner specified by the Steering Committee and the Advisory Committee.
- Submit periodic progress reports in a standard format, as agreed upon by the Steering Committee and the Advisory Committee
- Adhere to the NHGRI policies regarding intellectual property, data release and human subjects and other policies as might be established during the course of this activity
- Accept and implement the common guidelines and procedures approved by the Steering Committee
- Accept and participate in the cooperative nature of the group
- Attend Steering Committee meetings

The P.I. of a specialized sequencing center will:

- Determine experimental approaches, design protocols, set project milestones and conduct experiments
- If appropriate, produce genomic sequence to meet a quality standard and cost agreed upon at the time of award
- Release data according to NHGRI policies and publish results
- If appropriate, submit data for quality assessment by the quality control center or in any other manner specified by the Steering Committee and the Advisory Committee.
- Submit periodic progress reports in a standard format, as agreed upon by the Steering Committee and the Advisory Committee
- Adhere to the NHGRI policies regarding intellectual property, data release and human subjects and other policies as might be established during the course of this activity
- Accept and participate in the cooperative nature of the group
- Accept and implement the common guidelines and procedures approved by the Steering Committee
- Attend Steering Committee meetings

The P.I. of the quality control center will:

- In collaboration with the research network, determine experimental approaches, design protocols, and conduct quality assessment of the genomic sequence produced by the research network
- Release results of the quality assessment to NHGRI and back to each P.I.
- Submit periodic progress reports in a standard format, as agreed upon by the Steering Committee and the Advisory Committee
- Accept and participate in the cooperative nature of the group
- Accept and implement the common guidelines and procedures approved by the Steering Committee

- Attend Steering Committee meetings

3. NHGRI Program Staff Responsibilities:

The NHGRI Program Director(s) will have substantial scientific/programmatic involvement during the conduct of this activity through technical assistance, advice and coordination such as participating in the design of Research Network activities, advising in the selection of sources or resources, coordinating or participating in collection and/or analysis of data, advising in management and technical performance, or participating in the preparation of publications. However, the role of NHGRI will be to facilitate and not to direct the activities. It is anticipated that decisions in all activities will be reached by consensus of the Research Network and that NHGRI staff will be given the opportunity to offer input to this process. The NHGRI Program Director(s) shall participate as a member of the Steering Committee having one vote.

The Program Director(s) will:

- Participate (with the other Steering Committee members) in the group process setting research priorities, deciding optimal research approaches and protocol designs, and contributing to the adjustment of research protocols or approaches as warranted. The Program Director(s) will assist and facilitate the group process and not direct it.
- Serve as liaison, helping to coordinate activities among the awardees; act as a liaison to the NHGRI, and as an information resource about extramural genome research activities.
- Attend the Steering Committee meetings as a voting member, assist in developing operating guidelines, quality control procedures, and consistent policies for dealing with recurrent situations that require coordinated action. The Program Director(s) must be informed of all major interactions of members of the Steering Committee. The NHGRI Program Director(s) will be responsible for scheduling the time and preparing concise (3 to 4 pages) minutes or a summary of the Steering Committee meetings, which will be delivered to members of the group within 30 days after each meeting.
- Lend his/her relevant expertise and overall knowledge of the NHGRI- and NIH-sponsored research to facilitate the selection of scientists not affiliated with the awardee institutions who are to serve on the Advisory Committee and the Steering Committee.
- Serve as liaison between the Steering Committee and the Advisory Committee, attending Advisory Committee meetings in a non-voting liaison member role.
- Serve on subcommittees of the Steering Committee and the Advisory Committee, as appropriate.
- Provide advice in the management and technical performance of the investigation.
- The Program Director(s) will serve as scientific liaison between the awardees and other program staff at NHGRI.
- Assist in promoting the availability of the human genome sequence and related resources to the scientific community at large.
- Retain the option to recommend the withholding or reduction of support from any project within the Research Network that substantially fails to achieve its sequencing goals at the

quality stated in the NHGRI sequence quality standard and at a cost agreed at the time the goals are set for the next year, fails to remain state of the art in its production sequencing capabilities, fails to release data according to the Terms and Conditions of the award, or fails to comply with any other term of the award.

- o Participate in data analyses, interpretations, and where warranted, co-authorship of the publication of results of studies conducted through the Research Network.

4. Collaborative Responsibilities

The Steering Committee will serve as the main governing board of the Research Network. The Steering Committee membership will include the NHGRI Program Director(s), the P.I. from each awarded cooperative agreement (including those of the production centers, the specialized sequencing centers and the quality control center), and three research scientists with relevant expertise, but who are not affiliated with any of the cooperative agreements. The rest of the steering committee will appoint these three members by majority vote. One of these three members will be nominated to serve as the Chair of the Steering Committee and will be appointed by the Program Director(s). Additional members may be added by action of the Steering Committee. Other government staff may attend the Steering Committee meetings, if their expertise is required for specific discussions.

The Steering Committee will be responsible for discussing progress within the Research Network, and for advising NHGRI as to how the Research Network can complete the human DNA sequence within the stated goals of time and accuracy, and within budget. The Steering Committee will work with the quality control center to develop uniform procedures for data quality assessment. Members of the Steering Committee will be required to accept and implement the common guidelines and procedures approved by the Steering Committee.

Within one month after award of the cooperative agreements, the NHGRI Program Director(s) and the P.I.s will meet (perhaps by telephone conference) to select the three outside committee members and to nominate a chair from among those three. The Program Director(s) will appoint the Chair and schedule the first meeting of the Steering Committee once the Chair has been selected. The Chair of the Steering Committee will be responsible for coordinating the Committee's activities, preparing meeting agendas, and chairing meetings. A meeting schedule will be developed at the first meeting. Two meetings will be held each year, either in Bethesda or at one of the sites. One of the meetings will partially overlap with the annual meeting of the Advisory Committee. The purpose of meeting jointly will be to allow direct interaction between members of the Research Network and the Advisory Committee, prior to the latter's annual evaluation of the Research Network's progress. Subcommittees will be established by the Steering Committee as it deems appropriate.

5. Advisory Committee

The Advisory Committee will be responsible for reviewing and evaluating the progress of the Research Network toward completing that portion of the human DNA sequence for which

NHGRI is responsible. The Advisory Committee will be composed of four to six senior scientists with relevant expertise. The Director, NHGRI, will select the members and Chair. The membership of the Advisory Committee may be enlarged permanently, or on an *ad hoc* basis as needed.

The Advisory Committee will meet at least once a year. The first part of this meeting will be a joint meeting with the Steering Committee to allow the Advisory Committee members to interact directly with the members of the Research Network. Annually, the Advisory Committee will make recommendations regarding progress of the Research Network and present advice about changes which may be necessary in the Research Network program to the Director, NHGRI.

6. Arbitration Process

Any disagreement that may arise on scientific/programmatic matters (within the scope of the award), between award recipients and the NHGRI may be brought to arbitration. An Arbitration Panel, composed of three members - one Research Network Steering Committee designee, one NHGRI designee, and a third designee with expertise in the relevant area and chosen by the other two designees, will be convened. This special arbitration procedure in no way affects the awardee's right to appeal an adverse action that is otherwise appealable in accordance with PHS regulations 42 CFR Part 50, Subpart D and HHS regulation at 45 CFR Part 16.

7. Yearly Milestones

Awardees will be asked to define yearly milestones at the time of the award and to adjust these milestones annually at the anniversary date. In accord with the procedures described above, NHGRI may withhold or reduce funds for projects that substantially fail to meet their milestones or to maintain the center at the state of the art.

INCLUSION OF WOMEN AND MINORITIES IN RESEARCH INVOLVING HUMAN SUBJECTS

It is the policy of the NIH that women and members of minority groups and their subpopulations must be included in all NIH supported biomedical and behavioral research projects involving human subjects, unless a clear and compelling rationale and justification is provided that inclusion is inappropriate with respect to the health of the subjects or the purpose of the research. This new policy results from the NIH Revitalization Act of 1993 (Section 492B of Public Law 103-43) and supersedes and strengthens the previous policies (Concerning the Inclusion of Women in Study Populations, and Concerning the Inclusion of Minorities in Study Populations), which have been in effect since 1990. The new policy contains some provisions that are substantially different from the 1990 policies.

All investigators proposing research involving human subjects should read the "NIH Guidelines For Inclusion of Women and Minorities as Subjects in Clinical Research," which have been published in the Federal Register of March 9, 1994 (FR 59 14508-14513) and reprinted in the

NIH Guide for Grants and Contracts, Volume 23, Number 11, March 18, 1994. Investigators also may obtain copies of the policy from the program staff listed under INQUIRIES. Program staff may also provide additional relevant information concerning the policy.

LETTER OF INTENT

Prospective applicants are asked to submit, by August 1, 1998, a letter of intent that includes a descriptive title of the overall proposed research, the name, address and telephone number of the Principal Investigator, the number and title of this RFA, and a list of the key investigators and their institution(s) and projects. Any applicant planning to submit an application for more than $500,000 direct cost in any one year must contact the NHGRI staff listed under the INQUIRIES section in order for the application to be accepted by NIH.

The letter of intent should be sent to:
Dr. Jane L. Peterson
Program Director, Large Scale Sequencing
National Human Genome Research Institute
National Institutes of Health
38 Library Drive, MSC 6050
Building 38A, Room 614
Bethesda, MD 20892-6050

PUBLIC BRIEFING ON THE RESEARCH NETWORK FOR LARGE-SCALE SEQUENCING OF THE HUMAN GENOME

Propsective applicants are invited to attend a briefing on this Research Network program on May 13, 1998 in the Plimpton Room of the Beckman Center at the Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. NHGRI staff will explain the purpose of the program, provide detailed instructions about the application process and answer questions. Applicant institutions are urged to send a representative to this briefing. For further information about the meeting or accommodations in the area, please contact the program staff listed in this RFA.

APPLICATION PROCEDURES

The research grant application form PHS 398 (rev. 5/95) (see web site: http://www.nih.gov/grants/funding/phs398/phs398.html) is to be used in applying for these grants. Application kits are available at most institutional offices of sponsored research and may be obtained from the Division of Extramural Outreach and Information Resources, National Institutes of Health, 6701 Rockledge Drive, MSC 7910, Bethesda, MD 20892-7910, telephone [redacted] and from the program administrator listed under INQUIRIES.

The RFA label available in the PHS 398 (rev. 5/95) application form must be affixed to the bottom of the face page of the application. Failure to use this label could result in delayed

processing of the application such that it may not reach the review committee in time for review. In addition, the RFA title and number must be typed on line 2 of the face page of the application form and the YES box must be marked.

Submit a signed, typewritten original of the application, including the Checklist, and three signed photocopies, in one package to:

CENTER FOR SCIENTIFIC REVIEW
NATIONAL INSTITUTES OF HEALTH
6701 ROCKLEDGE DRIVE, SUITE 1040 - MSC 7710
BETHESDA, MD 20892-7710
BETHESDA, MD 20817 (for express/courier service)

At the time of submission, two additional copies of the application must also be sent to:

Dr. Rudy Pozzatti
Scientific Review Administrator
Office of Scientific Review
National Human Genome Research Institute
National Institutes of Health
38 Library Drive, MSC 6050
Building 38A, Room 609
Bethesda, MD 20982-6050

[redacted]

Applications must be received by October 9, 1998. If an application is received after that date, it will be returned to the applicant without review. The Center for Scientific Review (CSR) will not accept any application in response to this RFA that is essentially the same as one currently pending initial review, unless the applicant withdraws the pending application. The CSR will not accept any application that is essentially the same as one already reviewed.

REVIEW CONSIDERATIONS

A. General Considerations

Upon receipt, applications will be reviewed for completeness by CSR and responsiveness by the NHGRI. Incomplete applications will be returned to the applicant without further consideration. If NHGRI staff find that the application is not responsive to the RFA, it will be returned without further consideration.

Applications that are complete and responsive to the RFA will be evaluated for scientific and technical merit by an appropriate peer review group convened by the NHGRI in accordance with the review criteria stated below. As part of the initial merit review, a process (triage) may be

used by the initial review group in which applications will be determined to be competitive or non-competitive based on their scientific merit relative to other applications received in response to the RFA. All applications will receive a scientific review and summary statement, although applications judged to be competitive will be discussed and be assigned a priority score. Applications determined to be non-competitive will be withdrawn from further consideration and the principal investigator/program director and the official signing for the applicant organization will be promptly notified. The second level of review will be provided by the National Advisory Council for Human Genome Research.

All applications will be judged on the basis of the scientific and technical merit of the proposed projects and the documented ability of the investigators to meet the RESEARCH OBJECTIVES of the RFA.

B. Review Criteria

The application must be directed toward attaining the programmatic goals as stated under RESEARCH OBJECTIVES. The following criteria will be used by peer review groups to evaluate these applications:

For sequence production centers:

1. Likelihood that the project will produce a significant fraction of the complete human sequence:
   - o prior demonstrated success and quality of the proposed plan for a) producing high quality sequence and b) increasing throughput, including both upstream map production and sequence finishing
   - o prior demonstrated success and quality of the proposed plan for decreasing cost, including efficiency improvements due to technology development or other factors
   - o prior demonstrated success and quality of the proposed plan for identifying and solving critical integration problems, including adequacy of the informatics activities
2. Contribution of technology development:
   - o Success in incorporating new technologies, with an emphasis on how this has increased productivity and reduced cost, and the merit of the plans for incorporating additional new technologies; the promise of the proposed program of incorporation of new technologies to contribute to sequence production; and evidence that the center has maintained the 'state-of-the-art' in sequencing technology
3. Sequence quality:
   - o Merit of sequence quality assessment plans, including validation of fidelity to the genome, monitoring and minimizing sequencing errors, and other QA/QC plans
   - o Results from NHGRI sequence quality assessment exercises
   - o History of attaining—and proposed measures to improve—overall contiguity, including increasing the length of, and minimizing gaps (including N's) in, the finished sequence; this includes contiguity within and between clones

4. Track Record of the P.I. and other key personnel

5. Quality of the management plan, including workflow, scale-up, divisions of labor/responsibility among components, coordination between components, appropriate staffing, training, *etc.*
6. Past compliance with NHGRI data release policies, and plans for data release
7. Availability of the facilities, resources, expertise and technology necessary to perform the research, and the level of institutional commitment
8. Appropriateness of the proposed budget and time-line in relation to the proposed research

For the specialized sequencing projects:

1. Likelihood that the project will contribute to the completion of the first sequence of the Human Genome
2. Value, significance or unique role of the proposed research in contributing to the overall sequencing effort, both as an independent project and as a part of the overall sequencing effort undertaken by the other participants in the Research Network described here
3. Quality of the plans to integrate any new technology development into a large-scale sequencing effort
4. Track record of the P.I. and other key personnel
5. Past compliance and plans for data release
6. Availability of the facilities, resources, expertise and technology necessary to perform the research, and the level of institutional commitment
7. Appropriateness of the proposed budget and time-line in relation to the proposed research

For the quality control center:

1. Quality of the plan to assess the accuracy, contiguity and fidelity of the DNA sequence being produced by the production sequencing centers or where appropriate, the specialized sequencing projects.
2. Quality of the plans to provide assistance to the projects within the Research Network
3. Track Record of the P.I. and other key personnel
4. Quality of the management plan
5. Availability of the facilities, resources, expertise and technology necessary to perform the research, and the level of institutional commitment
6. Appropriateness of the proposed budget and time-line in relation to the proposed quality assessment program

The second level review will be conducted by the National Advisory Council for Human Genome Research.

AWARD CRITERIA

Awards will be made on the basis of scientific and technical merit as determined by peer review, including the significance of the projected contribution toward meeting the NHGRI program goal of contributing to the completion of the human DNA sequence by the year 2005, program needs

and balance, adherence to NHGRI policies on human subjects, data release and intellectual property, and the availability of funds.

INQUIRIES

Written and telephone inquiries concerning this RFA are encouraged. The opportunity to clarify issues or questions about the RFA from potential applicants is welcome.

Direct inquiries regarding programmatic issues to:

Dr. Jane L. Peterson
Dr. Adam Felsenfeld
Division of Extramural Research
National Human Genome Research Institute
National Institutes of Health
38 Library Drive, MSC 6050
Building 38A, Room 614
Bethesda, MD 20892-6050
[redacted]

Direct inquiries regarding fiscal matters to:

Ms. Jean Cahill
Grants Management Office
National Human Genome Research Institute
Building 38A, Room 613,
38 Library Drive, MSC 6050
Bethesda, MD 20892-6050
[redacted]

Schedule

Public Briefing at Cold Spring Harbor, NY: May 13, 1998
Letter of Intent Receipt Date: July 1, 1998
Application Receipt Date: October 9, 1998
Scientific Review Date: Feb/March 1999
Advisory Council Date: May 1999

Anticipated Award Date: July 1999

AUTHORITY AND REGULATIONS

This program is described in the catalog of Federal Domestic Assistance No. 93.172. Awards are made under the authority of the Public Health Service Act, Title IV, Part A (Public Law 78-410, as amended by Public Law 99-158, 42 USC 241 and 285) and administered under PHS grants policies and Federal Regulations 42 CFR Part 52 and 45 CFR Parts 74 and 92. This program is not subject to the intergovernmental review requirements of Executive Order 122372 or Health Systems Agency review.

OUTLINE OF THIRD 5 YEAR PLAN FOR THE HGP **DRAFT 1-7-98**

This outline summarizes recommendations developed at the various workshops to date. It is intended to help organize the discussion, but not to endorse all the enumerated goals.

I. Introduction

This section will describe the purpose of the plan, how it was arrived at, the NIH-DOE collaboration, etc

II. Background on HGP and previous plans

This section will describe briefly the origins, history and overall mission of the HGP.

III. Current goals and status

This section will describe progress to date in terms of the world-wide status, with no attempt to tease out individual contributions.

Should progress be reported goal by goal as in the current 5year plan? Or highlights summarized something like the following?

A.	genetic map	finished in 1994	
B.	physical map	genome-wide YAC/STS map done, goal for STSs met, some chromosomes have much more detailed maps,	
C.	sequence	___ bp of human DNA sequenced, using criteria agreed to	
D.	gene identification	technology developed, e.g. arrays ___human ESTs sequenced,___mapped	
E.	model organisms	E. coli	done
		Yeast	done
		C. elegans	complete ?
		Drosophila	___% completed
		mouse	genetic map done ___STSs mapped ___ESTs sequenced, mapped ___ bp of mouse sequence done

F. ELSI to be supplied by ERPEG

IV. New Goals

Should these be the NIH/DOE goals or world-wide goals? We have been making goals for NIH/DOE in the past as if we were the only players, but reporting progress on a world-wide basis.

A. DNA Sequencing

1. Human

a. Complete the human DNA sequence by 2005 *(define what this means, heterochromatin?)*

Quality standards: Aim for 99.99% accuracy
Aim for contiguity over at least 500,000 bases with irreducible gaps annotated as to size and orientation
Include confidence levels for each base

b. Build up a sustained sequencing capacity that will allow continued high throughput, low cost sequencing of the genomes of additional organisms, even beyond the completion of the human genome sequence.

2. Model Organisms

a. Drosophila complete by 2002

b. Mouse Continue to sequence syntenic regions
ESTs, sequence and map
Bac map???
Sequence???
(to be completed after workshop)

c. Other Identify other model organisms that can make major contributions to understanding of the human genome and support appropriate genomic studies
(discuss criteria in the accompanying text)

3. Technology

a. Improve throughput and reduce cost of state of the art sequencing technology

b. Develop new technology that will allow the sequencing of one complex genome per year at affordable cost.

B. Sequence Variation

1. Create an initial resource of DNA samples and cell lines for use in polymorphism studies with representation of individuals whose ancestors derive from diverse geographic areas.

2. Explore the need for additional population resources

3. Develop technology for rapid identification and scoring of SNPs

4. Create a SNP map of at least 100,000 SNPs. *(Discuss desirability of SNPs from coding regions in text)*

5. Eventually identify all common polymorphisms in known genes and catalogue all common haplotypes in human DNA

C. Functional Analysis

1. Sequence the full inserts in a representative set of human ESTs

2. Develop a database of expression patterns for human and model organisms, including internal standards to allow cross-comparison *(is this a genome goal or broader than that?)*

3. Support the development of technology for areas such as:
   obtaining full-length cDNAs
   finding rare transcripts
   large-scale in situ analysis
   high-throughput cis-element analysis
   identifying a complete set of protein folds
   large-scale protein expression analysis
   comprehensive protein interaction analysis

D. Bioinformatics

1. Improve integration and utility of databases
2. Develop better methods for analyzing sequence homology and variation
3. Develop efficient methods for whole genome association studies
4. Develop methods for large-scale haplotype analysis and linkage disequilibrium studies

5. Develop ways of representing comprehensive expression and function data electronically
6. Develop new analytical tools for expression and function data
7. Develop tools for displaying data such as maps visually

E. ELSI

To be supplied by ERPEG

F. Training

Nurture multipliscinary training, especially in bioinformatics

NOTE: Blanks in the progress report section will be filled in as of the date of submission for publication.

The goals will be amplified by explanatory text that also mentions caveats, constraints etc, similar to the 1993 plan.

There will also be general philosophical statements about sharing of data, availability of materials, public databases, emphasis on technology development, etc.

Reprint Series
1 October 1993, Volume 262, pp. 43–46

SCIENCE

A New Five-Year Plan for the U.S. Human Genome Project

Francis Collins and David Galas*

A New Five-Year Plan for the U.S. Human Genome Project

Francis Collins and David Galas*

The U.S. Human Genome Project is part of an international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms. Thanks to advances in technology and a tightly focused effort, the project is on track with respect to its initial 5-year goals. Because 3 years have elapsed since these goals were set, and because a much more sophisticated and detailed understanding of what needs to be done and how to do it is now available, the goals have been refined and extended to cover the first 8 years (through September 1998) of the 15-year genome initiative.

In 1990, the Human Genome programs of the National Institutes of Health (NIH) and the Department of Energy (DOE) developed a joint research plan with specific goals for the first 5 years [fiscal year (FY) 1991–95] of the U.S. Human Genome Project (*1*). It has served as a valuable guide for both the research community and the agencies' administrative staff in developing and executing the genome project and assessing its progress for the past 3 years. Great strides have been made toward the achievement of the initial set of goals, particularly with respect to constructing detailed human genetic maps, improving physical maps of the human genome and the genomes of certain model organisms, developing improved technology for DNA sequencing and information handling, and defining the most urgent set of ethical, legal, and social issues associated with the acquisition and use of large amounts of genetic information.

Progress toward achieving the first set of goals for the genome project appears to be on schedule or, in some instances, even ahead of schedule. Furthermore, technological improvements that could not have been anticipated in 1990 have in some areas changed the scope of the project and allowed more ambitious approaches. Earlier this year, it was therefore decided to update and extend the initial goals to address the scope of genome research beyond the completion of the original 5-year plan. A major purpose of revising the plan is to inform and provide a new guide to all participants in the genome project about the project's goals. To obtain the advice needed to develop the extended goals, NIH and DOE held a series of meetings with a large number of scientists and other interested scholars and representatives of the public, including many who previously had not been direct participants in the genome project. Reports of all these meetings are available from the Office of Communications of the National Center for Human Genome Research (NCHGR) and the Human Genome Management Information System of DOE (*2*, *3*). Finally, a group of representative advisors from NIH and DOE drafted a set of new, extended goals for presentation to the National Advisory Council for Human Genome Research of NIH and the Health and Environmental Research Advisory Committee of DOE. These bodies have approved this document as a statement of their advice to the two agencies, and the following represents the goals for FYs 1994–98 (1 October 1993 to 30 September 1998).

General Principles

Several general observations underlie the specific goals (Fig. 1) described here. The first observation is that successful development of new technology for genomic and genetic research has been essential to the achievements of the project to date and will continue to be critical in the future. It was clearly recognized, both in the 1988 National Research Council (NRC) report (*4*) and in the first NIH-DOE plan, that attainment of the ambitious goals originally set for the genome project would require significant technological advances in all areas, such as mapping, sequencing, informatics, and gene identification. As the genome project has proceeded, progress along a broad range of technological fronts has been conspicuous. Among the most notable of these developments have been (i) new types of genetic markers, such as microsatellites, that can be assayed by polymerase chain reaction (PCR); (ii) improved vector systems for cloning large DNA fragments and better experimental strategies and computational methods for assembling those clones into large, overlapping sets (contigs) that compose useful physical maps; (iii) the definition of the sequence tagged site (STS) (*5*) as a common unit of physical mapping; and (iv) improved technology and automation for DNA sequencing. Further substantial improvements in technology are needed in all areas of genome research, especially in DNA sequencing, if the project is to stay on schedule and meet the demanding goals that are being set.

A second general observation concerns an evolution in the levels of biological organization at which genomic research will likely function over the next few years. Initially, attention was focused on the chromosome as the basic unit of genome analysis. Large-scale mapping efforts, in particular, were directed at the construction of chromosome maps. The sophisticated genetic linkage maps now available and the detailed physical maps that are being produced are clear measures of the success of that approach. However, other units of study for the Human Genome Project will also have increasing usefulness in the future. Therefore, further mapping efforts directed at both larger and smaller targets should be encouraged. At one end of the scale, "whole genome" mapping efforts, in which the entire genome is efficiently analyzed, have become feasible with developments in PCR applications and robotics. These approaches generally produce relatively low-resolution maps with current technology. At the other end of the scale, increasing attention needs to be paid to detailed mapping, sequencing, and annotation of regions on the order of one to a few megabases in size. Although small in comparison with the whole genome, a megabase is still large in comparison with the capabilities of conventional molecular genetic analysis. Thus, development of efficient technology for approaching detailed analysis of several-megabase sections of the genome will provide a useful bridge between conventional genetics and genomics, and provide a foundation for innovation from which future methods for analysis of larger regions may arise.

Third, a goal for identifying genes within maps and sequences, implicit in the original plan, has now been made explicit. The progress already made on the original goals, combined with promising new approaches to gene identification, allow this element of genome analysis to be given greater visibility. This increased emphasis on gene identification will greatly enrich the maps that are produced.

It must also be noted that, as in the original 5-year plan, these goals assume a funding level for the U.S. Human Genome Project of $200 million annually, adjusted for inflation. As the detailed cost analysis for the first 5-year plan was performed in

F. Collins is the director of the National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892.
D. Galas was associate director, Office of Health and Environmental Research, Department of Energy, Washington, DC 20585.

* Present address: Darwin Molecular, 2405 Carillon Point, Kirkland, WA 98033.

1991, a cost of living increase must be added for all years beyond FY 1991. This funding level has not yet been achieved (Table 1).

International Aspects

The Human Genome Project is truly international in scope, as the original planners envisioned it. Its success to date has been possible because of major contributions from many countries and the extensive sharing of information and resources. It is hoped and anticipated that this spirit of international cooperation and sharing will continue. This coordination has been achieved largely by scientist-to-scientist interaction, facilitated by the Human Genome Organization (HUGO), which has taken on responsibility for some aspects of the management of the international chromosome workshops in particular. These workshops have served to encourage collaboration and the sharing of information and resources and to facilitate the expeditious completion of chromosome maps.

Several notable individual international collaborations have marked the genome project so far. One is the United States–United Kingdom collaboration on the sequencing of the *Caenorhabditis elegans* genome. Scientists at the Los Alamos National Laboratory are collaborating with Australian colleagues to develop a physical map of chromosome 16, and investigators at the Lawrence Livermore National Laboratory are working with Japanese scientists on a high-resolution physical map of chromosome 21. Other joint efforts include the collaboration between NIH and the Centre d'Etudo du Polymorphisme Humain (CEPH) on the genetic map of the human genome and the Whitehead/Massachusetts Institute of Technology–Généthon collaboration on the whole-genome approach to the human physical map. These are but examples of the myriad interrelationships that have formed, generally spontaneously, among participating scientists.

Specific Goals

Genetic map. The 2- to 5-cM human genetic map of highly informative markers called for in the original goals is expected to be completed on time. However, improvements to make the map more useful and accessible will still be needed. If the field develops as predicted, there will be an increasing demand for technology that allows the nonexpert to type families rapidly for medical research purposes. In addition, to study complex genetic diseases, there is a need to be able to easily test large numbers of individuals for many markers simultaneously. In the long run, polymorphic markers that can be screened in a more automated fashion, and methods of gene mapping that obviate the need for a standard set of polymorphic markers are also desirable.

Table 1. The budget for the Human Genome Project for NIH and DOE (in millions of dollars). Budgets for 1994 and 1995 have not yet been determined.

Fiscal year	NIH	DOE	Total	1991 Projection of Needs
1991	87.4	47.4	134.8	135.1
1992	104.8	61.4	166.2	169.2
1993	106.1	64.5	170.6	218.9
1994				246.8
1995				259.9

Goals

(i) Complete the 2- to 5-cM map by 1995.
(ii) Develop technology for rapid genotyping.
(iii) Develop markers that are easier to use.
(iv) Develop new mapping technologies.

Physical map. An STS-based physical map of the human genome is expected to be available in the next 2 to 3 years, with some areas mapped in more detail than others and an average interval between markers of about 300 kb. However, such a map will not likely be sufficiently detailed to provide a substrate for sequencing or to be optimally useful to scientists searching for disease genes. The original goal of a physical map with STS markers at intervals of 100 kb remains realistic and useful and would serve both sequencers and mappers. Using widely available methods, a molecular biologist can isolate a gene that is within 100 kb of a mapped marker, and a sequencer can use such a map as the basis for preparing the DNA for sequencing. To the extent that they do not introduce statistical bias, the use of STSs with added value (such as those derived from polymorphic markers or genes) is encouraged because such markers add to the usefulness of the map.

Goal

(i) Complete an STS map of the human genome at a resolution of 100 kb.

Physical maps of greater than 100-kb resolution are needed for DNA sequencing, for the purpose of finding genes and for other biological purposes. Although a variety of options are being explored for creating such maps, the optimal approach is by no means clear. There is a need to develop new strategies for high-resolution physical mapping as well as new cloning systems that are well integrated with advanced sequencing technology. Technology for sequencing is evolving rapidly. Therefore, preparation of sequence-ready sets of clones should be closely associated with an imminent intent to sequence.

There is a pressing need for clone libraries with improved stability and lower chimerism and other artifacts and a need for better technology for traveling from one STS to the next. A greater accessibility to clone libraries should also be encouraged.

DNA sequencing. Although the goal of sequencing DNA at a cost of $0.50 per base pair may be met by 1996 as originally projected, the rate at which DNA can be sequenced will not be sufficient for sequencing the whole human genome. Priority should be given during the next 5 years to increasing sequencing capacity by increasing the number of groups oriented toward large-scale production sequencing. Substantial new technology that will allow sequencing at higher rates and lower costs is also needed: evolutionary technology developed from improvements in current gel-based approaches and revolutionary technology developed on the basis of new principles. These developments will only occur if significantly greater financial resources can be invested in this area. It is estimated that an immediate investment of $100 million per year will be needed for sequencing technology alone, to allow the human genome to be sequenced by the year 2005.

Goals

(i) Develop efficient approaches to sequencing one- to several-megabase regions of DNA of high biological interest.
(ii) Develop technology for high throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
(iii) Build up a sequencing capacity to a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

The standard model organisms should be sequenced as rapidly as possible, with *Escherichia coli* and *Saccharomyces cerevisiae* completed by 1998 or earlier and *C. elegans* nearing completion by 1998. It is often advantageous to sequence the corresponding regions of human and mouse DNA side by side in areas of high biological interest. The sequencing of full-length, mapped complementary DNA molecules is useful, especially if it is associated with technological innovation applicable to genomic sequencing.

The measurement of the cost of sequencing is complex and fraught with many uncertainties due to the diversity of approaches being used. However, we need to continue to reduce costs, as well as im-

prove our ability to assess the accuracy of the sequence produced. This latter point must be addressed in future sequencing efforts. Cost will be highly dependent on the level of accuracy achieved.

Gene identification. Identification of all the genes in the human genome and in the genomes of certain model organisms is an implicit part of the Human Genome Project. Although the previous 5-year plan did not explicitly identify this activity with a specific goal, progress in mapping and in technology now makes it desirable to do so. With both genetic and physical maps of the human genome and the genomes of certain model organisms becoming available and large amounts of sequence data beginning to appear, it is important to develop better methods for identifying all the genes and incorporating all known genes onto the physical maps and the DNA sequences that are produced. This information will make the maps most useful to scientists studying the involvement of genes in health and disease. While many promising approaches are being explored, more development is needed in this area.

Goals

(i) Develop efficient methods of identifying genes and for placement of known genes on physical maps or sequenced DNA.

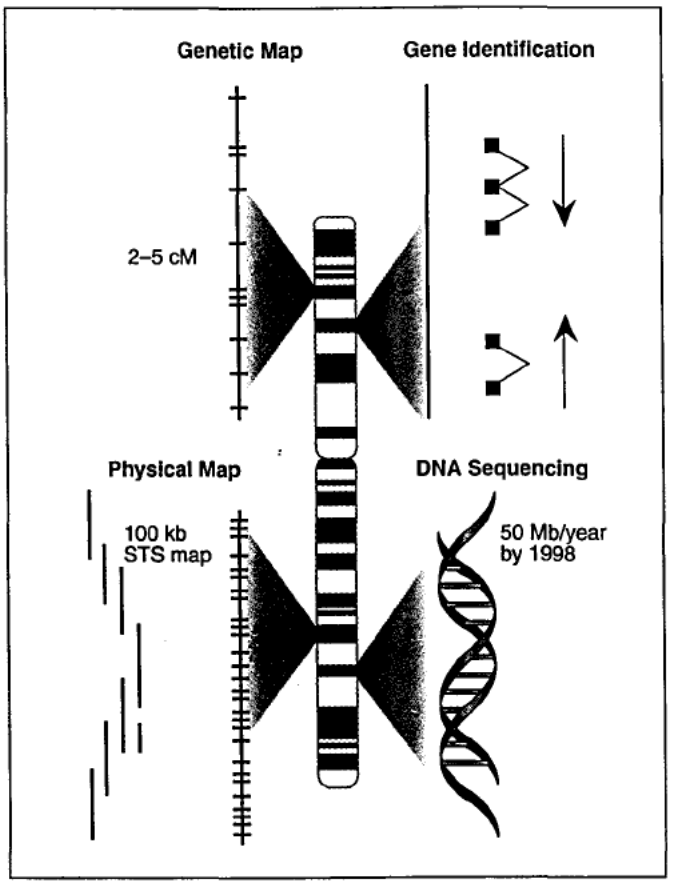


Fig. 1. Graphic overview of the new goals for the human genome. A 2- to 5-cM genetic map is expected to be completed by 1995 and a physical map with STS markers every 100 kb by 1998. Efficient methods for gene identification need to be developed and refined. The DNA sequencing goal of 50 Mb per year by 1998 includes all DNA, both human and model organisms, and assumes an exponential increase in sequencing capacity over time. Other important goals involving model organisms are not shown here, but are described in the text.

Technology development. The development of new and improved technology is vital to the genome project. Certain technologies, such as automation and robotics, cut across many areas of genome research and need particular attention. Cooperation in technology development should be encouraged where possible because it is likely to be more effective and efficient than competition and duplication. The technology developed must be expandable and exportable, the long-term goal being to create technology that will be available in many basic science laboratories and allow the efficient sequencing of other genomes. Technology development is costly and has not been sufficiently funded.

Goal

(i) Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and to meet the needs of the Human Genome Project as a whole.

Model organisms. Excellent progress has been made on the mouse genetic map and the *Drosophila* physical map, as well as the sequencing of the DNA of *E. coli*, *S. cerevisiae*, and *C. elegans*. Many of the original goals for this area are likely to be exceeded. Completion of the mouse map and sequencing of all the selected model organism genomes continue to be high priorities. The current emphasis for sequencing of mouse DNA should be placed on the sequencing of selected regions of high biologic interest side by side with the corresponding human DNA.

Goals

(i) Finish an STS map of the mouse genome at 300-kb resolution.

(ii) Finish the sequence of the *E. coli* and *S. cerevisiae* genomes by 1998 or earlier.

(iii) Continue sequencing *C. elegans* and *Drosophila* genomes with the aim of bringing *C. elegans* to near completion by 1998.

(iv) Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest.

Informatics. In order to collect, organize, and interpret the large amounts of complex mapping and sequencing data produced by the Human Genome Project, appropriate algorithms, software, database tools, and operational infrastructure are required. The success of the genome project will depend, in large part, on the ease with which biologists can gain access to and use the information produced. Although considerable progress has been made in this area since the beginning of the genome project, there is a continuing need for improvements to stay current with evolving requirements. As the amount of information increases, the demand for it and the need for convenient access increase also. Thus, data management, data analysis, and data distribution remain major goals for the future.

Goals

(i) Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.

(ii) Consolidate, distribute, and continue to develop effective software for large-scale genome projects.

(iii) Continue to develop tools for comparing and interpreting genome information.

Ethical, legal, and social implications (ELSI). The ELSI components of the Human Genome programs of NIH and DOE are strongly connected with genomic research so that policy discussions and recommendations are couched in the reality of the science. To date, the focus of the ELSI programs has been on the most immediate potential applications in society of genome research. Four areas were identified by advisers to the ELSI program for initial emphasis: privacy of genetic information, safe and effective introduction of genetic information in the clinical setting, fairness in the use of genetic information, and professional and public education. The program gives strong emphasis to understanding the ethnic, cultural, social, and psychological influences that must inform policy development and service delivery. Initial policy options for genetic family studies, clinical genetic services, and health care coverage have been developed, and reports on a range of urgent issues are expected by 1995.

As the genome project progresses, the need to prepare for even broader public impact becomes increasingly important. Poli-

cies are needed to anticipate the potential consequences of widespread use of genetic tests for common conditions, such as genetic predisposition to certain cancers or genetic susceptibility to certain environmental agents. In addition, as the genetic elements of behavioral and other nondisease-related traits are better understood, increased educational efforts will be needed to prevent stigmatization or discrimination on the basis of these traits. Continued emphasis on public and professional education at all levels will be critical to achieving these goals. Mechanisms for developing policy options that build on the current research portfolio and actively involve the public, the relevant professions, and the scientific community need to be developed.

Goals

(i) Continue to identify and define issues and develop policy options to address them.
(ii) Develop and disseminate policy options regarding genetic testing services with potential widespread use.
(iii) Foster greater acceptance of human genetic variation.
(iv) Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues.

Training. There is a continuing need for individuals highly trained in the interdisciplinary sciences related to genome research. The original goal of suporting 600 trainees per year proved to be unattainable, because the capacity to train so many individuals in interdisciplinary sciences did not exist. However, now that a number of genome centers have been established, it is anticipated that training programs will expand. Although no numerical goal is specified, expansion of training activities should be encouraged, provided standards are kept high. Quality is more important than quantity.

Goal

(i) Continue to encourage training of scientists in interdisciplinary sciences related to genome research.

Technology transfer. Technology transfer is already occurring to a remarkable extent, as evidenced by the number of genome-related companies that are forming. Many interactions and collaborations have been established between genome researchers and the private sector. In addition to the need to transfer technology out of centers of genome research, there is also a need to increase the transfer of technology from other fields into the genome centers. Increased cooperation with industry, as well as continued cooperation between the agencies, is highly desirable. Care must be taken, however, to avoid conflicts of interest.

Goal

(i) Encourage and enhance technology transfer both into and out of centers of genome research.

Outreach. It is essential to the success of the Human Genome Project that the products of genome research be made available to the community. However, only a subset of the total information is likely to be of interest at any one time, with the nature of that subset changing over time. Therefore, it is desirable to have flexible distribution systems that respond quickly to user demand. The private sector is best suited to this situation and has begun to play an active and highly valued role. This should be encouraged and facilitated where possible, including the provision of seed funding in some instances.

The NIH and DOE genome programs have adopted a rule for sharing of information: Newly developed data and materials are to be released within 6 months of their creation. This policy has been well accepted. In many instances, information has been released before the end of the 6 months.

Goals

(i) Cooperate with those who would establish distribution centers for genome materials.
(ii) Share all information and materials within 6 months of their development. The latter should be accomplished by submission of information to public databases or repositories, or both, where appropriate.

Conclusion

To date, the Human Genome Project has experienced gratifying success. However, enormous challenges remain. The technology that will lead to the sequencing of the entire human genome at reasonable cost must still be developed. Major support of research in this area is essential if the genome project is to succeed in the long run. The new goals described here are designed to address the long- and short-term needs of the project.

Although there is still debate about the need to sequence the entire genome, it is now more widely recognized that the DNA sequence will reveal a wealth of biological information that could not be obtained in other ways. The sequence so far obtained from model organisms has demonstrated the existence of a large number of genes not previously suspected. For example, almost half of the open reading frames identified in the genomic DNA of *C. elegans* appear to represent previously unidentified genes. Similar results have been observed in both *S. cerevisiae* and *E. coli* genomic DNA. Comparative sequence analysis has also confirmed the high degree of homology between genes across species. It is clear that sequence information represents a rich source for future investigation. Thus, the Human Genome Project must continue to pursue its original goal, namely, to obtain the complete human DNA sequence. At the same time, it is necessary to assure that technologies are developed that will allow the full interpretation of the DNA sequence once it is available. In order to increase emphasis on this area, an explicit goal related to gene identification has been added.

The genome project has already had a profound impact on biomedical research, as evidenced by the isolation of a number of genes associated with important diseases, such as Huntington's disease, amyotrophic lateral sclerosis, neurofibromatosis types 1 and 2, myotonic dystrophy, and fragile X syndrome. Genes that confer a predisposition to common diseases such as breast cancer, colon cancer, hypertension, diabetes, and Alzheimer's disease have also been localized to specific chromosomal regions. All these discoveries benefitted from the information, resources, and technologies developed by human genome research. As the genome project proceeds, many more exciting developments are expected including technology for studying the health effects of environmental agents; the ability to decipher the genomes of many other organisms, including countless microbes important to agriculture and the environment; as well as the identification of many more genes involved in disease. The technology and data produced by the genome project will provide a strong stimulus to broad areas of biological research and biotechnology. Exciting years lie ahead as the Human Genome Project moves toward its second set of 5-year goals.

REFERENCES AND NOTES

1. U.S. Department of Health and Human Services and Department of Energy, *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years* (April 1990).
2. National Institutes of Health, National Center for Human Genome Research, Office of Communications, Bethesda, MD 20892. Phone, (301)402-0911; Fax, (301)402-4570.
3. U.S. Department of Energy, Human Genome Management Information System, Oak Ridge National Laboratory, PO Box 20008, Oak Ridge, TN 37831-6050. Phone, (615) 576-6669; Fax, (615) 574-9188.
4. National Research Council, Committee on Mapping and Sequencing the Human Genome, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
5. M.V. Olson, L. Hood, C. Cantor, D. Botstein, *Science* **245**, 143 (1989).

HUMAN GENOME PROJECT

Physicists Urge Technology Push to Reach 2005 Target

In the 1980s, the Department of Energy (DOE) was the first U.S. agency to invest in the Human Genome Project, an attempt to decipher the human genetic code. But DOE's role has been overshadowed in the 1990s by a well-funded latecomer, the National Institutes of Health (NIH). In 1998, for example, NIH will spend roughly $218 million on its National Institute for Human Genome Research. DOE, in comparison, will spend $87 million on genome work. But both agencies have pledged to support the same objective—to determine all 3 billion bases in the human genome by 2005.

This target seemed ambitious when DOE and NIH adopted it 5 years ago. And for DOE, which recently acquired a new management team and is revamping its program, it still looks very difficult. Indeed, this winter, DOE has received a new warning about the difficulty of the task from a group of advisers known as the JASONs—an independent group of physicists and engineers who got together in 1960 to advise the military on weapons design. In a report issued this month, they warn that unless DOE and NIH make a significant improvement in technology used to sequence the genome, they may not reach their goal by 2005.

DOE seems to be taking this warning seriously, although it lacks the budget to make the sizable investment in technology that the JASONs recommend. NIH's genome institute director, Francis Collins, seems less concerned. "We are confident," he says, that with the ramp-up in genetic sequencing being promised by NIH-funded centers, "we will reach a level of output ... that will allow us to cross the finish line in good form." Collins says that if a current snapshot of production is correct, NIH-funded centers will generate 80 million bases of human genomic sequence this year, and will reach an annual production rate of 400 million to 500 million bases in a few years. DOE's program has contributed less than 2 million bases thus far, but hopes to scale up to 20 million in 1998 (see chart).

Seeking advice. Patrinos asked JASONs to look at genome project.

The JASONs were invited to take a look at DOE's contribution to the effort by Ari Patrinos, the mechanical engineer who took charge of DOE's human genome portfolio last year. For more than a year, Patrinos has been trying to hitch up three big genome research groups at the Los Alamos, Lawrence Livermore, and Lawrence Berkeley National Laboratories to a common strategic plan, and he looked to the JASONs for help.

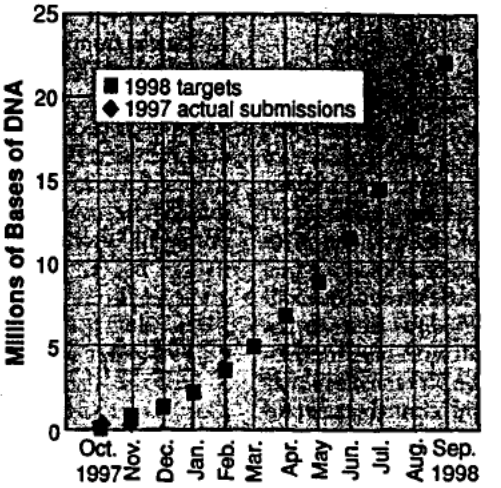
Patrinos says he had worked with the JASONs on climate change in the past and thought they might provide a fresh view of the genome project from "outside the community." A group headed by computational physicist Steven Koonin, a provost of the California Institute of Technology in Pasadena, received a briefing from top genome scientists last spring and toured the three major sequencing centers. This month, they released the first of what Patrinos expects will be several advisory reports (see p. 36).

In their report, the JASONs suggest that DOE's top priority should be to develop "advanced technology" for futuristic genome research, and they urge DOE to increase its budget for such research by 50% (to $20 million a year). The report also makes several suggestions for improving current technology, such as creating a "user group" to improve the performance of gel electrophoresis machines. In addition, it says that DOE should develop a quantitative approach to assessing the quality of genomic data, drawing upon the weapons labs' expertise in computational analysis. DOE should also improve the management of its own genomic databases, the JASONs concluded.

Arguing that sequencers desperately need more efficient machines to read DNA and better software to analyze the output, the JASONs report warns that "if the DOE does not continue to play a leading role in technology development ... it is not clear to us who will." Collins takes issue with this remark, which he considers "not fully informed." He says NIH spends more than $22 million a year on technology development and at least $2.5 million on ideas that have "nothing to do with" current technology.

At DOE, Patrinos must decide how to fit the advice into what he calls "a major reform" of DOE's genome program, which is now under way. Sequencing teams based at three DOE laboratories have been consolidated into a single Joint Genome Initiative, headed by DOE bioinformatics expert Elbert Branscomb, formerly of Livermore in California. Meanwhile, DOE has leased two buildings to create a new "sequencing factory" in Walnut Creek, California. Patrinos has told DOE's genome staff that its first job is to get the factory running and crank out 20 million bases of human genomic DNA within the next year—which he agrees is a "tough" objective. DOE soon will solicit bids to fill its new Walnut Creek factory this summer with $6 million worth of new sequencing equipment.

DOE's SEQUENCING GOALS



SOURCE: DOE

Tough targets. DOE has set ambitious goals for this year's sequencing effort.

In some areas, the DOE reform dovetails with the JASONs' advice. The report insists, for example, that "quality issues must be brought to the fore," and it proposes that DOE fund new research on ways to make sure that the published sequences are accurate. Patrinos notes that he is planning a new set of DNA quality-control standards, along with a major overhaul of database management—although not all the details have been disclosed. The report also recommends a "systems approach" to mass-producing biological data and suggests that lab managers create "error budgets" for each stage in the production process. These ideas, which some DOE scientists have called "naïve," are under review.

But Patrinos acknowledges that there is "some tension" between his decision to emphasize a rapid ramp-up in the output of sequence data and the JASONs' recommendation that DOE focus on new technology. He concedes that, without a larger budget, the DOE project cannot invest as much as he would like in new technology projects. For now, he sees no alternative but to emphasize production, noting that he's "a little nervous" about meeting the 2005 target.

–Eliot Marshall

POLICY: BIOMEDICINE

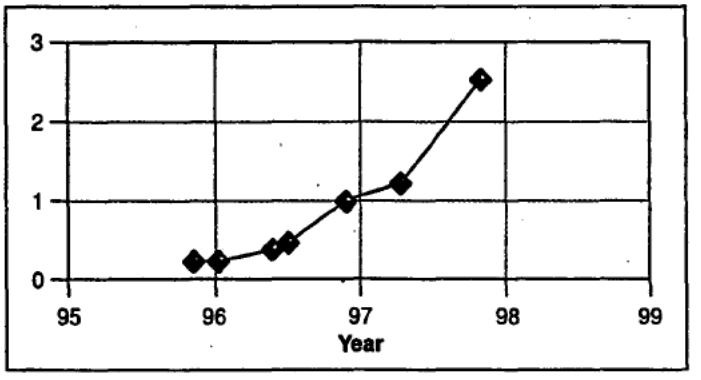
An Independent Perspective on the Human Genome Project

Steven E. Koonin

The U.S. Human Genome Project (HGP) is a joint effort of the Department of Energy and the National Institutes of Health, formally initiated in 1990. Its stated goal is ". . . to characterize all the human genetic material—the genome—by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining the complete sequence . . . to discover all of the more than 50,000 human genes and render them accessible for further biological study." The original 5-year plan was updated and modified in 1993 (*1*, *2*).

DOE's Office of Biological and Environmental Sciences recently chartered the JASON group to review the DOE component of the HGP. This group, mainly consisting of physical and information scientists, was asked to consider three areas: technology, quality assurance and quality control, and informatics. This article summarizes the group's findings and recommendations (*3*).

Technology. The present state of the art for determining the sequence of DNA is defined by Sanger sequencing, in which DNA fragments are labeled by fluorescent dyes and separated according to length with polyacrylamide gel electrophoresis (PAGE) (*4*). The base at the end of each fragment can then be visualized and identified by the dye with which it reacts. Although more than 95% of the genome remains to be sequenced, roughly 55 megabases (Mb) have been completed in the past year (see the figure). The world's large-scale sequencing capacity (not all of which is applied to the human genome) is estimated to be roughly 100 Mb per year. It is sobering to contemplate that an average production of 400 Mb will be required each year to complete the human sequence by the target date of 2005.



Percentage of the human genome sequenced to date. Almost 3% of the genome has been sequenced in contiguous stretches longer than 10 kb and is now deposited in publicly accessible databases. Compiled by J. Roach, as described in http://weber.u.washington.edu/~roach/.human_genome_progress2.html.

The present technology has only a limited read-length capability (the number of contiguous bases that can be identified from each fragment); the best current practice can read 700 to 800 bases, with perhaps 1000 bases as the ultimate limit. Because the DNA segments of interest are much longer than this [40 kilobases (kb) for a cosmid clone; 100 kb or more for a bacterial artificial chromosome or a gene], the present technology requires that long lengths of DNA be cut into overlapping short segments (~1 kb in length) that can be sequenced directly. The sequences from these shorter pieces must then be assembled into the final sequence. Up to 50% of the effort at some sequence centers goes into this final assembly and finishing of the sequence. The ability to read longer fragments would step up the pace and quality of sequencing.

Apart from the various genome projects, however, there is little pressure to achieve longer read lengths. The 500 to 700 base lengths read by the current technology are well suited to many scientific needs, including pharmaceutical searches, studies of some polymorphisms, and studies of some genetic diseases.

Other drawbacks of the present technology include the time- and labor-intensive nature of gel preparation and running, as well as the comparatively large amounts of sample required, which also increases the cost of reagents and necessitates extra amplification steps.

Thus, the present sequencing technology leaves much to be desired and must be supplanted in the long term if the potential for genomic science is to be fully realized. Promising methods that could be cheaper and faster than PAGE include single-molecule sequencing, mass spectrometric methods, hybridization arrays, and microfluidic capabilities. None of these is sufficiently mature, however, to be a candidate for near-term major scale-up. It is therefore important to support research aimed at improving the present method. Advances in hardware development could, for example, increase the lateral scan resolution of the machine so that more lanes of a gel can be analyzed. The genome community should unify its efforts to enhance the performance of present-day instruments.

Better software will improve the lane tracking, base identification, assembly, and finishing processes. Many of the problems of base identification also occur in the demodulation of signals in communication and magnetic recording systems, and some of the existing literature in these areas should be used by the HGP. The ability to correctly assemble a final sequence without manual editing would markedly speed up the process. It would also be helpful to develop a common set of finishing rules.

Because sequencing technology should (and is likely to) evolve rapidly, the large-scale sequencing centers must be flexible enough to incorporate new technologies. There is a great need to support the development of non-PAGE–based sequencing that goes beyond the current goals of a faster version of PAGE. The funding for such advanced technology is a small fraction of the total HGP but should be increased by approximately 50%.

Quality assurance and quality control. DOE and NIH are recognizing that the HGP must make data accuracy and data quality integral to its execution. A high-quality database can provide useful, densely spaced markers across the genome and enable large-scale statistical studies. A quantitative understanding of data quality across the whole genome sequence is thus almost as important as the sequence itself. Among the top-level steps that should be taken are allocating resources specifically for quality issues and establishing a separate research program for quality assurance and control (perhaps a group at each sequencing center).

The author is professor of Theoretical Physics and vice president and provost at the California Institute of Technology. He led the JASON study reported on in this article. E-mail: koonin@caltech.edu

The stated accuracy goal of the HGP is one error in 10^4 bases, which is set to be less than the polymorphism rate. However, this has been a controversial issue, as genomic data of lower accuracy are still of great utility. For example, pharmaceutical companies searching for genes can use short sequences (400 bases) at an accuracy of one error per 100 bases. The debate on error rates should focus on the level of accuracy needed for each specific scientific objective or use of the genome data. The necessity of finishing sequences without gaps should be subject to the same considerations.

In the real world, accuracy requirements must be balanced against what users need, the cost, and the capability of the sequencing technology to deliver a given level of accuracy. Establishing this balance requires an open dialogue among the sequence producers, sequence users, and the funding agencies, informed by quantitative analyses and experience.

Assays should be developed that can accurately and efficiently measure sequence quality. For example, it would be appropriate to develop, distribute, and use "gold standard" DNA samples that could be used routinely by the whole sequencing community for assessing the quality of the sequence output.

Research into the origin and propagation of errors through the entire sequencing process is fully warranted. We see two useful outputs from such studies: (i) more reliable descriptions of expected error rates in final sequence data, as a companion to database entries; and (ii) "error budgets" to be assigned to different segments of mapping and sequencing processes to aid in developing the most cost-effective strategies for sequencing and other needs.

DOE and NIH should solicit and support detailed Monte Carlo computer simulation of the complete mapping and sequencing processes. The basic computing methods are straightforward: a reference segment of DNA (with all of the peculiarities of human sequence) is generated and subjected to models of all steps in the sequencing process; individual bases are randomly altered according to errors introduced at the various stages; and the final reconstructed segment or simulated database entry is compared with the input segment and errors are noted.

Results from simulations are only as good as the models used for introducing and propagating errors. For this reason, the computer models must be developed in close association with technical experts in all phases of the process being studied, so that they best reflect the real world. This exercise will stimulate new experiments to validate the error-process models and thus will lead to increased experimental understanding of process errors as well.

Improved software is needed to enhance the ability of database centers to check the quality of submitted sequence data before its inclusion in the database. Many of the current algorithms are highly experimental and will be improved substantially over the next 5 years. In addition, an ongoing software quality assurance program should be considered for the large community databases, with advice from commercial and academic experts on software engineering and quality control. It is appropriate for the HGP to insist on a consistent level of documentation, both in the published literature and in user manuals, of the methods and structures used in the database centers that it supports. DOE and NIH should also decide on standards for the inclusion of quality metrics for base identification and DNA assembly along with every database entry submitted.

Informatics. Genome informatics is a child of the information age, a status that brings clear advantages and new hurdles. Managing such a diverse, large-scale, rapidly moving informatics effort is a considerable challenge for both DOE and NIH. The infrastructure supporting the requisite software tools ranges from small research groups (for example, for local special-purpose databases) to large Genome Centers (for process management and robotic control systems) to community database centers (for GenBank and the Genome Database). The resources that each of these groups can put into increasing software sophistication, into ensuring ease of use, and into quality control vary widely. Thus, in informatics areas requiring new research (such as gene finding), a broad-based approach of "letting a thousand flowers bloom" is most appropriate. At the other end of the spectrum, DOE and NIH must impose community-wide standards for software consistency and quality in areas of informatics in which a large user community will be accessing major genome databases.

DOE and NIH should adhere to a bottom-up, customer approach to informatics. Part of this process would be to encourage forums, including close collaborative programs, between the users and providers of informatics tools, with the purposes of determining what tools are needed and of training researchers in the use of new methods.

To ensure that all the database centers are user-oriented and that they are providing services that are genuinely useful to the genome community, each database center should be required to establish its own "users group" (as is done by facilities as diverse as the National Science Foundation's Supercomputer Centers and NASA's Hubble Space Telescope). Further, informatics centers must be critically evaluated as to the actual use of their information and services by the community.

Data formats, software components, and nomenclature should be standardized across the community. If multiple formats exist, it would be worthwhile to invest in systems that can translate among them. Data archiving, data retrieval, and data manipulation should be modularized so that one database is not overextended, and several groups should be involved in the development effort. The community should be supporting several database efforts and promoting standardized interfaces and tools among those efforts.

Final notes. The HGP involves technology development, production sequencing, and sequence utilization. Greater coupling of these three areas can only improve the project. Technology development should be coordinated with the needs and problems of production sequencing, whereas sequence generation and informatics tools must address the needs of data users. Promotion of such coupling is an important role for the funding agencies.

The HGP presents an unprecedented set of organizational challenges for the biology community. Success will require setting objective and quantitative standards for sequencing costs (capital, labor, and operations) and sequencing output (error rate, continuity, and amount). It will also require coordinating the efforts of many laboratories of varying sizes supported by multiple funding sources in the United States and abroad.

A number of diverse scientific fields have successfully adapted to a "big science" mode of operation (nuclear and particle physics, space and planetary science, astronomy, and oceanography are among the prominent examples). Such transitions have not been easy on the scientists involved. However, in essentially all of these cases, the need to construct and allocate scarce facilities has been an important organizing factor. No such centralizing force is apparent in the genomics community, but the HGP is very much in need of the coordination it would produce.

References and Notes

1. F. Collins and D. Galas, *Science* **262**, 43 (1993).
2. The status and challenges of the HGP have been recently reviewed [L. Rowen *et al.*, *ibid.* **278**, 605 (1997)].
3. The MITRE Corporation, JASON Report JSR-97-315 (The MITRE Corporation, McLean, VA, 1997). The participants included S. Block, J. Cornwall, W. Dally, F. Dyson, N. Fortson, G. Joyce, H.J. Kimble, N. Lewis, C. Max, T. Prince, R. Schwitters, P. Weinberger, and W. H. Woodin.
4. For a basic discussion and explanation of the terminogy used, see http://www.ornl.gov/TechResources/Human_Genome/publicat/primer/intro.html

Draft 1/13/98

See pg 3 – response of subcomm to fen goals

OUTLINE OF THIRD 5 YEAR PLAN FOR THE HGP

I. Introduction

This section will describe the purpose of the plan, how it was arrived at, the NIH-DOE collaboration, etc

II. Background on HGP and previous plans

This section will describe what has been accomplished by the HGP to date

III. Current goals and status

The current goals will be enumerated and an accounting of progress on them given. Progress will be in terms of the world-wide status, with no attempt to tease out individual contributions.

A.	genetic map	finished in 1994	
B.	physical map	genome-wide YAC/STS map done, goal for STSs met, some chromosomes have much more detailed maps,	
C.	sequence	___ bp of human DNA sequenced, using criteria agreed to	
D.	gene identification	technology developed, e.g. arrays ___human ESTs sequenced,___mapped	
E.	model organisms	E. coli	done
		Yeast	done
		C. elegans	complete ?
		Drosophila	___% completed
		mouse	genetic map done ___STSs mapped ___ESTs sequenced, mapped ___ bp of mouse sequence done
F.	ELSI	to be supplied by ERPEG	

IV. New Goals

(Need to decide whether NIH/DOE or world wide)

A. DNA Sequencing

1. Human

a. Complete the human DNA sequence by 2005 (define what this means, heterochromatin?)

Quality standards: Aim for 99.99% accuracy
Aim for contiguity over at least 500,000 bases with irreducible gaps annotated as to size and orientation
Include confidence levels for each base

b. Build up a sustained sequencing capacity that will allow continued high throughput, low cost sequencing of the genomes of additional organisms, even beyond the completion of the human genome sequence.

2. Model Organisms

a. Drosophila complete by 2002

b. Mouse Continue to sequence syntenic regions
ESTs, sequence and map
Bac map???
Sequence???
(to be completed after workshop)

c. Other Identify other model organisms that can make major contributions to understanding of the human genome and support appropriate genomic studies
(discuss criteria in the accompanying text)

3. Technology

a. Improve throughput and reduce cost of state of the art sequencing technology
b. Develop new technology that will allow the sequencing of one complex genome per year at affordable cost.

B. Sequence Variation

1. Create an initial resource of DNA samples and cell lines for use in polymorphism studies with representation of individuals whose ancestors derive from diverse geographic areas.

2. Explore the need for additional population resources

3. Develop a SNP map of at least 100,000 SNPs. (Discuss desirability of SNPs from coding regions in text)

4. Eventually identify all common polymorphisms in known genes and catalogue all common haplotypes in human DNA

C. Functional Analysis

1. Sequence the full inserts in a representative set of human ESTs

2. Develop a database of expression patterns for human and model organisms, including internal standards to allow cross-comparison (is this a genome goal or broader than that?)

3. Support the development of technology for areas such as:
- ✓obtaining full-length cDNAs
- ✓finding rare transcripts
- large-scale in situ analysis
- high-throughput cis-element analysis
- +/- identifying a complete set of protein folds
- – large-scale protein expression analysis
- – comprehensive protein interaction analysis

D. Bioinformatics

1. Improve integration and utility of databases
2. Develop better methods for analyzing sequence homology and variation
3. Develop efficient methods for whole genome association studies
4. Develop methods for large-scale haplotype analysis and linkage disequilibrium studies
5. Develop ways of representing comprehensive expression and function data electronically
6. Develop new analytical tools for expression and function data
7. Develop tools for displaying data such as maps visually

E. Training

Nurture multipliscinary training, especially in bioinformatics

F. ELSI

To be supplied by ERPEG

NOTE: Blanks in the progress report section will be filled in as of the date of submission for publication.

The goals will be accompanied by explanatory text that also mentions caveats, constraints etc.

There will also be general philosophical statements about sharing of data, availability of materials, public databases etc.

AGENDA

Council Scientific Planning Subcommittee
January 13-14, 1998
Bethesda Marriott Hotel
5151 Pooks Hill Road
Bethesda, Maryland
Salon I
Congressional Ball Room

Tuesday, January 13, 1998 - 7:00 p.m. - 9:30 p.m.

Reports:

Resource Planning Workshop----------Lisa Brooks
Sequencing PI Meeting----------Jane Peterson
Function Workshop----------Elise Feingold

Other planned workshops:

Mouse----------Bettie Graham
Informatics/databases----------Lisa Brooks

Other items of interest:

DOE planning activities----------Marv Frazier
NIGMS workshops----------Chuck Langley, Lee Hartwell

Update on FY 1999 budget----------Elke Jordan

Wednesday, January 14, 1998 - 8:30 a.m. - 4:00 p.m.

Discussion of outline of 5-Year plan

Leroy Walters will join us around lunch time to report on ERPEG activities

*****We will work through lunch—food will be provided.**

By the end of the day, we should have a pretty clear idea of what the goals will look like for the areas where workshops have already taken place. The next 2-3 months will then be spent drafting the actual document, refining the goals, filling in the gaps, etc. No other meeting of the subcommittee is scheduled until May. We need to decide whether this is sufficient, i.e., can we work by e-mail and possibly conference call in the interim?

At the May 5/6 meeting we will need to integrate the DOE and NIH aspects into a draft that can be presented to the community at Airlie House on May 28/29.

AC 1/13/98

1998-2003 NHGRI PLANS and NOTES:

A) Genome Sequencing:

The next plan should remind and re-emphasize the major goal of the Human Genome Project — obtaining the reference genome DNA sequence of the human and model organisms. The emphasis on genomic sequence is not only to obtain the sequence of all genes but also that of controlling and regulatory elements of single genes, gene families, chromosomes and entire genomes. It is this that will enable us to transform biology and lead to new understanding of human disease, development and evolution. This will require DNA sequence data of high accuracy and long-range contiguity. There is every expectation that the DNA sequence can be "read" (computationally) in order to understand which are the coding versus regulatory sequences but to do so we require the reference genome sequence of a number of species across the evolutionary tree. To enable this paradigm to succeed we propose the following immediate 5 year goal:

Establish as a national resource the capacity for sustained DNA sequencing at the rate of 500 megabases (mb) per year and within 5 years.

Sequence accuracy: The standard for human DNA sequence accuracy is 99.99% using either two chemistries and/or by sequencing both strands. This rate has been determined based on the expectation that the probability of a true sequence difference between any two human DNAs is 0.1%; the sequencing "error" rate of 0.01% assures that the ovewhelming majority of differences encountered in the reference sequence will reflect valid polymorphic sites. Additionally, sequencing error leads to shorter predicted ORFs that may be difficult to reconcile with other experimental data. The cost of DNA sequencing is based on the set accuracy rate and is appropriate for the human. We have to discuss whether this is warranted for the mouse where a standard inbred strain can be sequenced. If the mouse is to be a serious candidate for genome sequencing then a somewhat reduced accuracy rate may be allowable and make this a more realistic goal.

Sequence contiguity: The challenge in genomic sequencing is to achieve long-range contiguity in the face of high accuracy. A recently accepted notion is to achieve average contiguity of 500 kilobases (kb), while reporting a frequency distribution of DNA sequence contig size. Human genome sequencers had previously accepted that an individual PI will sequence from one human linkage (Genethon) marker to the next, which, on average, is ~ 1 megabase (mb). We have to set standards regarding how the 500 kb average size should increase to 1,000 kb; perhaps at a rate of 100 kb/year or much more ?

Some of the difficulty in completing a contig is due to biological reasons, although some of the early experience has suggested that sequence errors can contribute to this in a major way. We need to set explicit rules as to what gaps in a sequence-ready contig can be tolerated at the DNA sequence level. This is an extremely important issue since difficult regions can cost 4 times more in "finishing" than easier regions.

Sequence cost: There are two views regarding how sequencing costs should be reported: either as $in/bases out or by per-lane accounting. Although there are specific advantages and disadvantages to either approach, for the subcommittee's purpose it appears that the $ in/bases out method can be simpler. For purposes of review of sequencing projects a per-lane accounting method appears more sensible since specific categories of cost can be identified and one can monitor which categories are expected to change on instituting changes in technology and procedures.

In either case, a "consensus" current estimate of large-scale genomic sequencing is $ 0.50/base. Our plan will crucially depend on what figures we shall assume. At the PI's meeting there was some agreement that a sequencing cost of $ 0.25/base could be achieved in 5 years with experience and scale-up, but that future reductions are hard to justify at this time, particularly if extensions to current technology are used. Thus, in my opinion, a cost reduction from $ 0.50 to $ 0.25 within 5 years may be feasible. Even then, some of the current sequencing groups may not be able to reduce cost while maintaining sequence quality and contiguity. The rate decrease corresponds to a per year decrease of 13% by all participating groups, and seems consistent with the 20-30% efficiency increase that some large-scale sequencing centers claim that they have achieved. Based on the PI presentations, a doubling of efficiency every 2 years appears difficult to maintain and, perhaps, untenable.

The cost of sequencing can vary widely with the local compositional "difficulty" of the sequence. A number of estimates suggested that the cost rate for "easy" versus "difficult" regions may be 1:4. It may well be that the human genome may have upto 30% of such difficult regions (read BAC clones) sequencing/finishing which may have a higher associated cost. We need some explicit rules on how this may be dealt with, although further sequencing experience may help greatly in this matter.

One of the difficulty of the current sequencing paradigm is the small amout of "finished" sequence obtained per lane: 50 bp at an accuracy of 0.01% with an average read length of 500 bp. Although there is much discussion on sequencing technology improvement, significant rapid benefits can be realized if

one could increase read length per lane and the number of reads per run. Although some of these gains can arise from changes in instrumentation, some anticipated by the sequencers and included in their plans for a cost reduction, a major benefit can arise only from further research into sequencing chemistry. Since sequencing technology and instrumentation is unlikely to change drastically over the next 5 years an emphasis on sequencing chemistry improvements is warranted.

Sequence target: The original stated aim of the HGP was obtaining the reference human genome sequence (RHGS) at 3 gb. Since other sequencing centers, not supported by the NHGRI/NIH, do exist, initial NHGRI plans showed that 60% of the sequence would be funded by the NIH. This plan needs to be revised for two reasons, a theoretical and a practical one. First, the importance of obtaining the genome sequence of an organism as a paradigm to do biology should be firmly established. In fact, I would argue that the genome sequence itself will reveal much of the interesting biology of that organism. This is particularly so given the power of comparative genomics which will require the genome sequence of a number of carefully chosen species. We should aim for this goal and attempt to make genomic sequencing as inexpensive as possible. Second, despite the plans of other organizations, it is both imperitive and important that NHGRI set the goal of sustainable sequencing capacity rather than the RHGS. The plans of other entities are not firm, not set on firm ground and divided over a number of sequencing agendas. It would be embarrasing if the set goal was lower than 3 gb by 2005 and other institutions failed to deliver (The Wellcome Trust has committed only 500 mb so far). This new goal, 500 mb per year within 5 years, will be ample to complete the RHGS even if some current centers fail to deliver, and will be particularly useful, if by unexpected gains in efficiency or by completion of the RHGS by contributions from DOE and the Wellcome Trust, is sequencing the mouse is contemplated.

My "radical" plan would be to concentrate on scale rather than on cost, in the near future, since it is that aspect (while maintaining accuracy and contiguity) which is the more difficult. I am assuming that the current cost of $ 0.50/bp can linearly decrease to $ 0.25 in 5 years by which time we need to develop a sustainable sequencing capacity of 500 mb per year. The costs can further decrease but its magnitude is not predictable. To enable this, I would recommend that we spend $ 80m in 1998 and increase it rapidly to $ 120m by 2000; a smaller increase to $ 125 from 2001 is also projected. This plan increases sequencing capacity by 25-40% over the next 5 years. A table describing this scenario is provided below. These figures show a ramp-up in sequencing capacity of 43%, 30%, 25% and 25% in years 2-5, respectively. This plan can allow the investigators to concentrate on decreasing sequencing cost while increasing capacity.

Year	Sequence Target (mb)	Cost rate ($/bp)	Total cost ($m)
1998	160	0.50	80
1999	229	0.4375	100
2000	320	0.3750	120
2001	400	0.3125	125
2002	500	0.25	125
2003	500	0.25	125
2004	500	0.25	125
2005	500	0.25	125

Sequence challenges:

The biological and medical benefits to obtaining the genome sequence of humans and other closely-related species are tremendous. However, to get there there are a number of significant hurdles that the sequencing community have to overcome. All of them revolve around the central requirement of achieving a large sustainable sequencing capacity at a small per base cost while maintaining accuracy and contiguity. Sequencing at this rate requires a size, style of organization and style of management that biologists are unaccustomed to and are having difficulty adapting to. There are not a large number of large-scale sequencing operations (no more than 10 in the USA) and not all of them may make the transition to larger sequencing centers. This may be due to their inability to argue for more space within their institution, their difficulty with recruiting/training/maintaining technical personnel or to achieve the stated scientific goals. Although it does appear that several sequencing groups currently exist most will not transit to centers producing 100 mb of genome sequence, and I believe that many will simply divert attention to other biological problems still involving sequencing. It is difficult to predict how many centers will exist and how much their output will be, and we do not know the failure rate. In fact, precisely to keep to the scientific goals, to keep enough sequencers occupied so that the best may emerge by competition, to reduce distractions and to allow for failures we need to infuse the process with more funds now. This is the principle of concentrating on scale rather than on cost. The Cooperative Agreement mechanism that NHGRI has discussed is an excellent way to

implement sequencing, vet out the successful groups, hold all groups to a common and high standard, as well as allow the improvements and experience in one center to be used by the others. What we all lack is the experience in truly large-scale sequencing and that is what we have to develop.

Sequence-ready maps:

There has been general consensus that sequence-ready clones for sequencing has not been a limiting factor. However, this view is not shared by all and the amount of mapping needed to support a 500 mb sequencing opertaion(s) may not exist within all centers. Indeed, this could become a rate limiting factor. Although a centralized mapping facility is not warranted at this time, NHGRI has to pay close attention to this problem. Furthermore, if the underlying mapping resources (clone libraries) are judged to be poor later this could also comp[romise the sequencing output.

GENERAL FUNDING GOALS:

1) The human and other genome sequence should drive other projects funded by the NHGRI. Newly funded grants must be able to utilize both the genome sequence and a genome-wide view.

2) All technology development must be geared to increasing quality, increasing throughput and decreasing cost, irrespective of whether this is for *de novo* sequencing, resequencing, expression studies or variation studies.

3) NHGRI may need a different subcommittee of the Advisory Council to assess whether NHGRI's funding portfolio is addressing the long-term needs of the scientific community and the Institute.

4) NHGRI needs to support both production genome sequencing and pioneer applications of the sequence. It is the latter that will allow the community, and NHGRI, to articulate our needs to the government, as well as guide us as to which applications are most fruitful.

5) Sequence-based biology needs to be funded with other NIH Institutes as partners, as NHGRI has demonstrated with the SNP project.

6) The HGP was to be funded at $ 200m/year. Admittedly, this was a crude guess in the NRC report but I am impressed by how close the figure is to what is currently needed. It is still not at that level; the FY1998 figure is at $171m sans the Intramural component which plays a major role in genome applications.

With now optimistic projections of a continuing large increase in the NIH budget over the next 5 years, it is appropriate to request specific increases. One scenario would be to request specific funding for technology development, particularly to develop sequencing technology that can produce a sustainable capacity of 1 gigabase (gb) of DNA sequence each year within 10 years at a cost of $0.05/base or better,i.e., an order of magnitude smaller than currently available. It is apparent that current methods cannot achieve this.

7) On a minor note, genome sequencing in Drosophila is currently at a cost of 50% greater than in the human. This distinction is not necessary and all production sequencing must have the same standards of cost efficiency and contiguity. We should also discuss the merits and disadvantages of sequencing at a accuracy of 99.99% when this figure has arisen from human polymorphism considerations. This may save some funds, but more importantly, will set the pace for all model organism sequencing.

8) NHGRI should lead the way in cost accounting for large biology projects, of which there will be many in the future. The principle of amortizing costs of fixed equipment etc. should be a recognizable feature of all grants, in particular, the sequencing agreement.

B) DNA Sequence Variation:

This area should be the second major goal in the next 5 year plan. The general goal should be the development of technology for discovery and scoring of all types of sequence-based DNA variants, in both the human and the mouse. A specific goal should be the generation of a set of 100,000 SNPs in the human.

C) Sequence-based functional analysis:

The major emphasis should be on the generation of expression maps in the mouse and human, probably in the former. We could make some real advances if we concentrated on expression studies as a function of developmental stages of specific tissues. Some funding for pilot projects on genome-wide protein studies, such as protein arrays, would be ideal.

D) Bioinformatics:

We have not had any discussion but need to; this has been an ommision on all of our part because we have been so preoccupied with the interesting stuff. However, the two major areas for support would be databases for genome sequence and its variants, and new databases for expression studies.

E) ELSI:

These will arise from the deliberations of ERPEG and will also be based on issues that the data and technologies in (A)-(D) raise.

F) Training:

We will need some focussed discussion on whom we wish to train, at what level and the areas where emphasis is needed. Clearly, training is like motherhood and apple-pie in that all of us are for it. However, in my view, without focus, funds allocated for this purpose may not find the best use. Our aims should be to concentrate on training individuals not so much on the technologies but rather on the emerging concept of a genome-wide view to biology, be that at the level of molecules, cells, tissues, individuals, families or populations.

Summary of Workshop on the Functional Analysis of Genomic Sequences

As part of NHGRI's five-year planning process, a workshop on the "Functional Analysis of Genomic Sequences" was held on December 2-3, 1997. The purposes of the workshop were to: (1) to define those biological questions which can be addressed using genomic approaches to gain insight into the biological function of genomic sequences, and (2) to explore the areas of new technology and resource development that will be required if genomic approaches to these questions are to be successful. The overall goal of the workshop was to develop a set of recommendations for areas/issues for NHGRI and the Council Subcommittee to consider in developing the next five-year plan, as well as for those areas/issues that are not reasonable to pursue further in the NHGRI planning process.

The two-day meeting began with six presentations that were personal visions of how biological research will evolve over the next 5-10 years, how genomics has already influenced the development of some fields, and what contributions genomic approaches could make in the future. The purpose of these talks was to set the stage for discussions rather than to present a comprehensive overview of the influence of genomics on biological research. Following these talks, there were three breakout sessions to discuss potential ideas for future genomic research. The participants were divided into three groups, one in each of the general areas of DNA analysis, RNA analysis, and protein analysis. Each group was moderated by a member of the National Advisory Council for Human Genome Research Planning Subcommittee. The participants were given a set of questions to facilitate discussion (see attached). The following day, a preliminary set of recommendations from each breakout group was reported by the moderator and discussed by the entire group of participants. In the final afternoon, these recommendations were refined into a more concise, non-redundant set.

What follows is a summary of the recommendations that were discussed at the final session and then a more detailed account of the points discussed in each breakout group.

DRAFT 1/6/98

Summary of Recommendations
Future NHGRI-Supported Research Efforts

Workshop on the Functional Analysis of Genomic Sequences
December 2-3, 1997

A. Production/Resources

1. Reference Human DNA Sequence

a. The completion of the sequence of the human genome was acknowledged to be of the highest priority for NHGRI.

2. Human SNPs

a. There was strong endorsement for NHGRI to pursue, in conjunction with other NIH Institutes, the generation of human SNPs as well as the development of tools to exploit them.

3. Full-Insert cDNA Sequences

a. There was consensus that these should be generated for the human; less consensus regarding the mouse (in part because of uncertainty as to what HHMI will support). An advantage of the mouse is that it will be possible to generate cDNA libraries with a different representation of genes than the human. Similar efforts for other model organisms, e.g. Drosophila, should be considered.

1% TX - w/NCI possible opp.

b. There was general consensus that one pass sequencing on each strand would provide adequate accuracy for human cDNAs, in part because it is anticipated that the genomic sequence will be done at a very high accuracy; accuracy for other organisms needs to be considered on a case-by-case basis. Confidence levels should be put on each base.

4. Other Model organisms

a. There is a need to establish criteria for determining whether or not to sequence any additional model organisms. A potential list of criteria was generated during the RNA session (see below), including "phylogenetic power," and the capability to transfect the organism. Consideration should be given to alternative approaches for some organisms (e.g. low pass or sequence-sampling strategies for genomic sequencing, or EST sequencing). In some instances, only the

Consensus - Seq Mouse

generation of genomic resources, such as genetic or physical maps, may be appropriate.

5. **Comprehensive "database" of RNA expression patterns in human and model systems**

   a. It would be valuable to create a database of RNA expression patterns that contains information about which sets of transcripts are expressed, and at what level, in each cell at any given stage of development, differentiation, or time in the cell cycle.

   b. There was general consensus that the technology for RNA expression analysis is sufficiently developed to initiate these types of projects now. However, there is a critical need for the development of internal standards to allow for the cross-comparison of studies. Additional technology development, especially in the area of informatics, is also needed (see RNA section below).

   c. This is a long-term goal (beyond the next 5 years), whose comprehensive achievement may be more appropriate for NIH as a whole than for NHGRI alone.

Microarray WS - much work - already in hated, some movement to getting technology to academic users.

B. Technology Development

Numerous opportunities for technology development were identified and recommended for support in the following areas:

- improving cDNA resources
- regulation of gene expression

1. **RNA Expression**

   a. Synthesis of full-length cDNA clones

      i. NHGRI's role in supporting the generation and sequencing of these cDNAs, once the technology has been robustly developed, needs further discussion.

   b. Discovery of rare/underrepresented transcripts

   c. Large-scale methods for RNA in situ analyses, including the development and use of multiple probes

   d. High-throughput cis-element analysis to study transcriptional regulation

   e. Defining regulatory hierarchies, such as the identification of all target genes regulated by a given factor or small combination of factors

quality / high information content

2. **DNA Analysis** - fcn of non-coding seq, esp those involved in chromosome / genome biology e.g. centromeres telomeres

   a. High-throughput analysis of non-coding sequences that function at the chromosomal level, such as centromeres and telomeres

3. **Protein Structure and Expression** fcn of proteins - thru structural analyses protein-expressing - where + when protein interactions

   a. Identification of the complete set of protein folds (thought to be finite in number, i.e., one to several thousand)

   b. Production of a complete set of expressed proteins

      i. Efficient methodology for heterologous expression of large quantities of proteins
      ii. Development of native protein microarrays

   c. Multiple, benign and readily recognizable protein tags for localization and other studies

   d. Large-scale protein expression analysis

      i. Improvement of 2D gels and other front end separation technologies for mass spectrometry
      ii. Improvement of mass spectrometry
      iii. Development of novel technologies, e.g. arrays of specific protein ligands

   e. Protein interactions

      i. Comprehensive analysis of protein-protein interactions, including protein complexes; further discussion of technology development for comprehensive analyses of protein-DNA and protein-ligand interactions as well as other physiological interactors is needed

C. Bioinformatics/Databases - not much discussion

1. New tools for data representation, visualization and analysis (including interactive/hierarchical data), e.g., computable pathway algorithms and electronic representation of metabolic pathways, are needed.

   capability to deal with complex sets of data

D. Training/Access

1. Computational biology training is critical.

2. There was less consensus regarding interdisciplinary training in other areas. One approach, thought by some to be more effective, is to build multidisciplinary research teams composed of individuals with specialized expertise and to nurture interdisciplinary collaborations.

3. Interdisciplinary training should be done at the post-Ph.D. level.

Comments: Although the participants endorsed sequencing of the mouse genome, there was no explicit discussion regarding this in the summary session. It was noted that there is going to be another workshop specifically focused on the mouse in March, 1998.

There were several other points that were strongly endorsed by one or more breakout groups that were not discussed at length in the summary session and might be considered for further discussion by the Council subcommittee. These include:

1. Facilitating/subsidizing affordable chip resources, access to genome technologies

2. Large-scale approaches to probe the function of gene products, e.g., mutagenesis/tagged insertions

K/o's / over ex pression

Summary of Recommendations by DNA Group
Maynard Olson

1. Reference "Databases"

- The generation of the first complete human genomic sequence was endorsed to be of the highest priority for NHGRI.

- The generation of a reference database for human polymorphisms was discussed at length. There was a strong consensus that NIH should be very active in this area, especially as it related to the generation of a large number of polymorphic markers (e.g. 100,000 SNPs), as well as additional theory development. A second, longer-term component (for which there was less consensus) was the comprehensive analysis of human polymorphisms. This type of analysis poses significant scientific as well as ELSI challenges and would require significant technology development.

- The sequencing of the mouse genome was not discussed at length, but should be considered for funding.

2. First-Pass Genome Resources

- Of overwhelming interest is the development of a strategy to obtain a relatively complete set of human cDNA sequences (and a similar resource for additional organisms if possible). This would not necessarily be a comprehensive set (including e.g. all splice variants and very rare transcripts) and may not need to be of highest accuracy nor from full-length clones, depending on the level of investment.

- There was somewhat less consensus on the development of additional first pass resources. These include EST sets for a number of organisms, beyond the standard models. A number of these sets would allow for better phylogenetic definition for higher organisms. Additional resources suggested were high-quality germline clone libraries and improved genetic maps for a variety of organisms.

- More research to study the function of germline sequences was endorsed by some of the members of the breakout group and this topic engendered significant discussion during the morning recap session, perhaps because of the strong opinions of a minority of the participants. Areas to pursue include the analysis of cis-regulatory regions controlling transcription and the functional analysis of other regulatory elements, such as those involved in chromosome structure, i.e. study the biology of the "genome" in addition to the genes. While there was considerable concern that this could be considered "the rest

of biology" some thought that genomic approaches to study these biological questions could be developed. One approach to support is mutagenesis, especially in the mouse. Further discussion is needed with respect to the relative merits of targeted (insertional/tagged) vs. chemical mutagenesis, and this topic will be addressed in the March, 1998 meeting on mouse genomic resources.

3. Technology Development

- There should be a major effort to push for a reduction in the cost of DNA sequencing. The genomes (or biologically interesting portions of genomes) of many model organisms could then be readily sequenced, which would alleviate the pressure to set strict priorities for choosing which additional model organisms (if any) to sequence. It was recognized that this is a very difficult problem requiring a significant investment. NHGRI should seek less traditional partners than have historically been considered (e.g., DARPA).

- Technology development for the generation of many of the first-pass resources discussed above is clearly needed.

4. Bioinformatics

- There was a significant level of enthusiasm for continued development in this area. It was recognized that there is a need for ongoing training at all levels and an emphasis on keeping a viable academic culture in this area. A vigorous small grants program is critically needed in this area to produce innovation and to maintain faculty in academia.

Additional Points Raised During the Discussion:

- NHGRI should take the lead in encouraging and facilitating the transfer of genomic resources to the general research community, not only from the large genome centers, but from individual labs as well.

- Promote the use of chips and other related technologies by increasing access and lowering the costs to researchers.

- It was stressed that there is significant value in sequencing model organisms beyond what will be learned about that given organism. If they are chosen in a phylogenetically-informed manner much can be learned about the human and other vertebrate organisms.

- The study of polymorphisms such as SNPs will also facilitate the functional analysis of the genome; some changes will be functionally significant.

Summary of Recommendations by RNA Group
Barbara Wold

1. Human and Mouse EST Resources

There was widespread enthusiasm for the current EST resources and further investment was thought to be highly worthwhile.

A. Resource Generation

- Validate the source of clones used to generate the existing human and mouse EST sets and complete the sequence of these clones. Validation would take approximately 6 months at an estimated cost of $1.5M, creating a higher quality resource that could be used for full-insert sequencing than currently exists.
- Construct an expression library for all existing full-length protein coding sequences
- Generate more full-length cDNAs

B. Technology Development

- Develop (and apply) new technologies for cloning underrepresented RNAs (low level expression; specific time and places)
- Improve expression vectors to allow for regulated expression in a variety of cell types and organisms

C. Other Considerations

- Encourage trans-NIH funding for resource generation
- Management and oversight of projects by NHGRI

2. Complete Molecular Phenotyping for Model Organisms

Determine what set of transcripts or proteins are expressed in each cell at a given time and at what level. This is a long-term goal (beyond the next 5 years) requiring significant technology development. Execution may go beyond NHGRI.

A. Technology Development

- Develop and implement internal standards for each model organism for inclusion in each data set for use in all methodological approaches. Will facilitate cross-comparisons
- Increase sensitivity of input with goal of single cell inputs

- Informatics to permit access; clear identifiers
- Informatics to link to different kinds of data
- Informatics/methods to assign a unique identifier, amount relative to standard and some kind of P value for this amount (analogous to quality standard for base calling) to each measurement
- Methods for cell enrichment
- Alternatives to array technology; alternate array technologies

B. Resource Generation

- Build standard data sets for expression studies for model organisms (continually update until complete array of genes)
- Provide "chips" (either complete set or subsets of genes) to user community at reasonable cost
- Provide technology access to R01 investigators
- Improve technology for export (cheaper, lower capacity if necessary)

C. Other Considerations

- Start with RNA first since technology is more advanced, then move to protein
- Challenge lies in determining site of resource generation: At center(s) vs. dissemination of technology

3. Characterizing "Wildtype" Mouse

Mouse phenotypes are poorly understood. Much underlying information is likely to have already been generated and there is a need to establish a means of capturing it in a central database.

A. Resource Generation

- Database of high quality phenotypic measurements (physiology, endocrinology, behavior, anatomy, etc) from standard strains used in knock-out experiments.

B. Other Considerations

- Combined informatics and new measurements
- Mandate R01 grantees doing knockout studies to submit wildtype data to "control" database
- Trans-NIH/other support
- Combined RFA/R01 contributions

4. Regulatory Architecture for Genome Expression

NHGRI should support technology development in this area; application of technology to specific areas may be more appropriately supported elsewhere.

- Develop (and apply?) technology to identify all target genes (functional cis-elements) regulated by a given factor or small combination of factors.

- Develop technologies for rapid, high-throughput cis-element discovery and characterization (couple biology and informatics)

- Develop methods for visual representation of complex, multidimensional, and often hierarchical data. There is a need for these methods to analyze many other types of large, complex data sets as well.

5. Additional Model Organisms

- Sequence the mouse genome

- Criteria for evaluation of candidates (to be used when sequencing costs come down)
  a) Transfection capability (essential)
  b) Phylogenetic power (essential)
  c) Mutagenesis/screening/strain maintenance
  d) Targeted mutagenesis (desirable)
  e) Availability of material, including embryos
  f) Genome size (preferably small)

- Possible candidate: Amphioxis or small genome tunicate prior to tetraploidy of vertebrates; avoid gene redundancy.

- Consider starting with EST projects for candidates; reduce pressure on genome size

6. Protein Structure/Manipulations

- High-throughput expression libraries for model organisms where you know all or most of the proteins (e.g. bacculovirus resource) followed by a massively parallel protein production and crystallization effort. Provide those that work to crystallography community

- Technology for improved crystallization methods designed to extend the range of proteins that can be handled. Support for the application of methods should be from resource interested in specific protein(s)

- Develop methods to render glycosylated proteins amenable for analysis by mass spectrometry

7. Additional New Technologies and Resources

These are clearly longer-term goals.

- Generate libraries of chemical ligands or antibodies for arraying, detecting, affinity purification of each protein for the model organisms and the human
- Develop technology (where still needed) for genome-wide, systematic (tagged) disruption of all genes in model organisms
- Generate resources of disrupted tagged strains as technology and finances permit. [Strain storage issues for some organisms]
- Methods for higher-order multiplexing of gene expression tags and in situ hybridization probes or protein detection probes (on the order of 10s –100s)

Additional Points Raised During the Discussion:

- While technology development is very important, the money required is beyond our budget. We need to consider partnerships with industry relatively early on in the development; exploit SBIR/STTR program; support proof of principle and then transfer it over to industry. There was some discussion about the implications of this approach, including access.

- Full-length cDNAs should be generated for all model organisms, or as many as possible.

Summary of Recommendations by Protein Group
Tony Pawson

1. General Recommendations (not related to proteins)

- Sequence mouse
- Improve the quality of the EST database
- Sequence full-length cDNAs (for predicting ORFs) from multiple organisms; complete accuracy not necessary

2. Protein Structure/Function

Work toward predicting function from protein sequence

- Understand totality of protein folds
    - Predict all possible folds
    - Analysis of novel folds by structural determination
- Improve homology modeling
- Improve alignments to assign protein families; take advantage of structural information
- Improve structural analysis of membrane proteins

3. Proteomics

Better technology is needed for quantitative global analysis of protein expression and post-translational modification.

- 2D gel technology
    - Improve technology for quantifying individual protein levels, identification of post-translational modifications. Needs standardization/automation/increased sensitivity. Useful currently for small genomes, further technology development needed for display of proteins from more complex systems.
    - Apply current technology to identify every protein in e.g., yeast/bacteria
- Mass Spectrometry
    - Technology development needed for front end (automation, sample loading/interfacing with separation technology) and back

end (software development, automated data collection and reference to databases)

- Protein Microarrays

    - Useful to identify protein ligands/physiological partners
    - Considered to be very important to develop, but highly challenging
    - Best done on domains
    - Should be group production effort using common technology; need to have specialists working with specific sets of proteins
    - Create analogous array of unique ligands to probe for protein expression
    - Develop novel methods for more rapid, automated technology for protein identification

4. Protein Interactions/Function

Generate set of reagents to allow you to learn about protein interactions and pathways

- Generate entire set of domains and identify peptide motifs (or other ligands) that they interact with; e.g., peptide libraries, phage display. Use to establish network of protein interactions.
- Generate similar set of affinity probes, e.g., small molecules or antibodies
- Develop global approaches to activate or inactivate protein
- Develop better prediction methods for protein localization
- Develop new technology to identify low affinity protein-protein & protein-ligand interactions

5. Bioinformatics

- Develop proteome database of higher eukaryotes serving as central organization of all that is known about proteins, e.g., motifs, structure, interactions, function.

6. Training

- Cross-discipline training important; suggested at the post-doctoral level rather than graduate student level.

Additional Points Raised During the Discussion:

- Strong endorsement of the approach to identify complete set of domains rather than the more brute force approach recommended by

the RNA group to determine the structure of every protein for which a crystal can be made; approach can be experimentally verified

- Suggested additional organism to sequence – one from the "bottom of the eukaryotic radiation." Many functions lost in yeast; study other unicellular organism.

A. Cost constant through FY2000, then decreases with a half life of 4 years

FY	$ for Production	$/bp	Mb produced
98	70	0.500	140
99	80	0.500	160
00	88	0.500	175
01	96.6	0.423	230
02	106	0.352	300
03	117	0.297	395
04	129	0.250	515
05	142	0.211	675
Total Sequence Produced			**2590**

B. Cost is constant through FY2000, then decreases 10% per year

FY	$ for Production	$/bp	Mb produced
98	70	0.500	140
99	80	0.500	160
00	88	0.500	175
01	97	0.450	215
02	106	0.405	260
03	117	0.364	320
04	129	0.328	395
05	142	0.295	480
Total Sequence Produced			**2145**

Cost decreases with a half-life of 4 years

FY	$ for Production	$/bp	Mb produced
98	60	0.500	120
99	62	0.423	147
00	64	0.352	182
01	66	0.297	222
02	68	0.250	272
03	70	0.211	332
04	72	0.176	409
05	74	0.149	497
Total Sequence Produced			**2181**

CONSTANT $ MODEL

	A	B	C	D	E	F	G	H
1	year	cost/base	years to go	capacity	cost	total sequence	total cost	cost/kb
2	1997	0.5	0	120	60	180	60	500
3	1998	0.45	1	(165) 147	66	327	126	450
4	1999	0.41	2	161	66	488	192	410
5	2000	0.37	3	178	66	666	258	370
6	2001	0.34	4	(960) 194	66	860	324	340
7	2002	0.31	5	213	66	1073	390	310
8	2003	0.29	6	228	66	1301	456	290
9	2004	0.27	7	244	66	1545	522	270
10	2005	0.25	8	264	66	1809	588	250

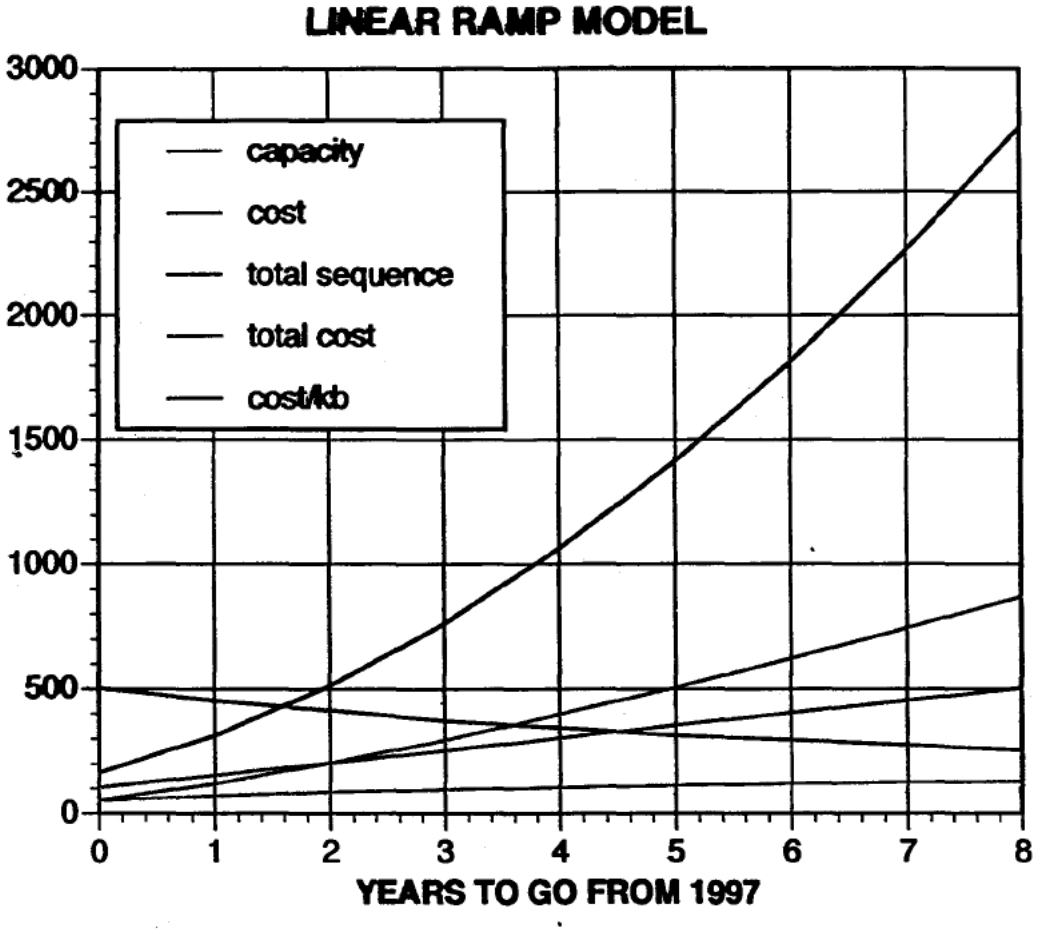
LINEAR RAMP MODEL

	A	B	C	D	E	F	G	H
1	year	cost/base	years to go	capacity	cost	total sequence	total cost	cost/kb
2	1997	0.5	0	115	58	175	58	500
3	1998	0.45	1	(165) 150	68	325	125	450
4	1999	0.41	2	200	82	525	207	410
5	2000	0.37	3	250	92	775	300	370
6	2001	0.34	4	(960) 300	102	1075	402	340
7	2002	0.31	5	350	108	1425	510	310
8	2003	0.29	6	400	116	1825	626	290
9	2004	0.27	7	450	122	2275	748	270
10	2005	0.25	8	500	125	2775	872	250

FAST RAMP MODEL

	A	B	C	D	E	F	G	H
1	year	cost/base	years to go	capacity	cost	total sequence	total cost	cost/kb
2	1997	0.5	0	115	58	175	58	500
3	1998	0.45	1	(165) 165	74	340	132	450
4	1999	0.41	2	230	94	570	226	410
5	2000	0.37	3	295	109	865	335	370
6	2001	0.34	4	(960) 360	122	1225	458	340
7	2002	0.31	5	465	144	1690	602	310
8	2003	0.29	6	500	145	2190	747	290
9	2004	0.27	7	500	135	2690	882	270
10	2005	0.25	8	500	125	3190	1007	250

( ) = PT



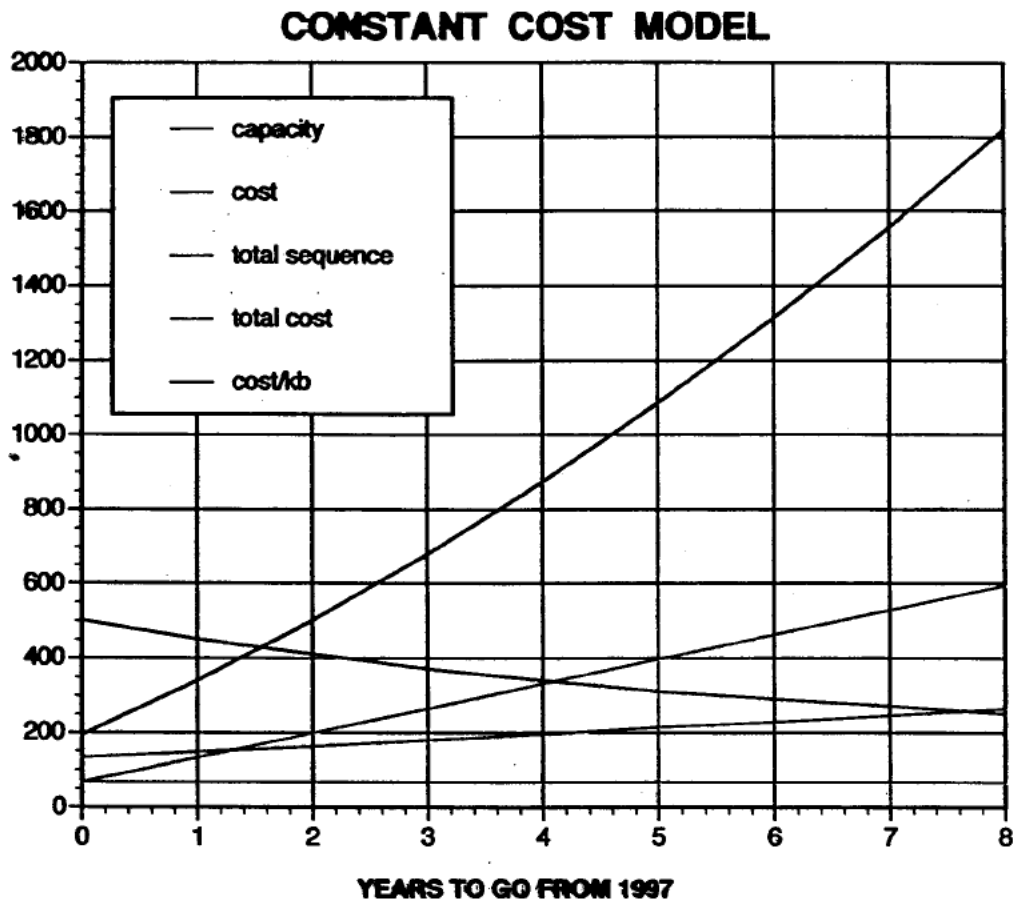
NOTES:

1. This smooth ramp in capacity (50 Mb/year) and hence $ is conservatively realistic and will allow for more than 2-3 centers.

2. This model provides ~1 Gb of excess capacity to be used for other organisms or as an error margin.

3. Continuing to ramp production towards the end is more cost effective, more psychologically reassuring, and easier to budget.

4. Other investments are not precluded.



NOTES:

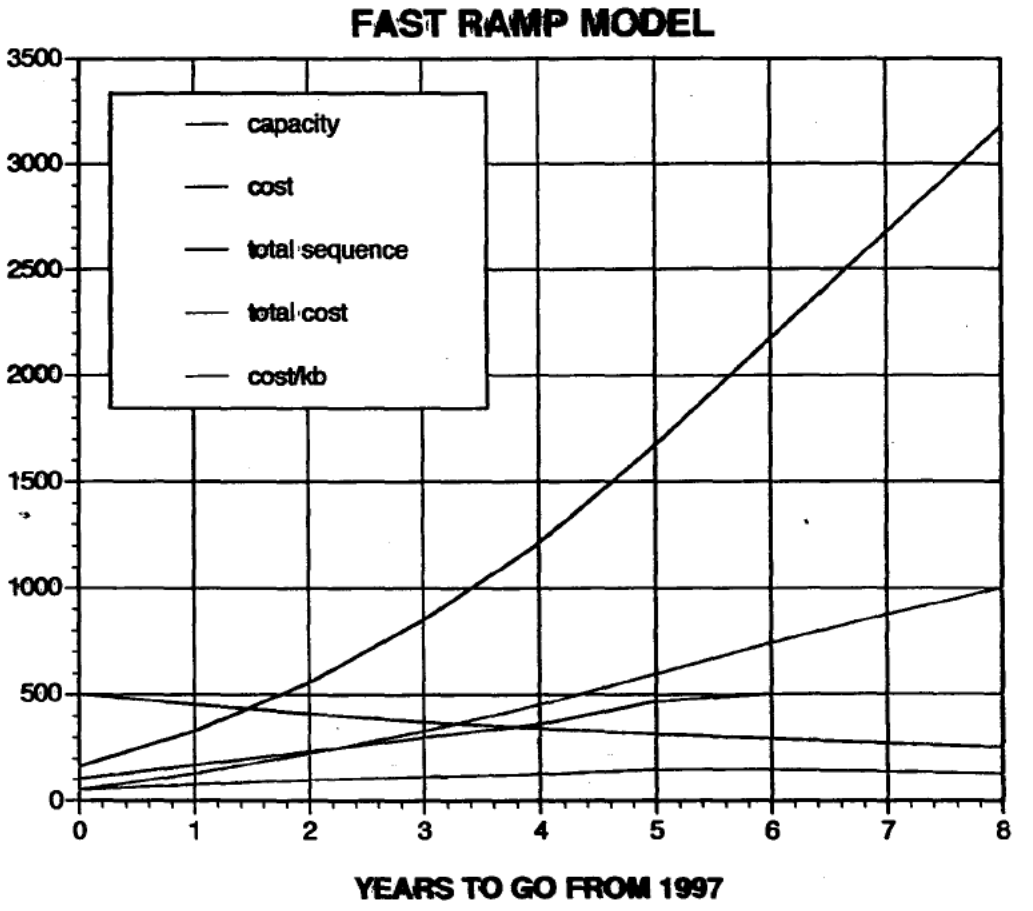
1. This model has no margin for error and just barely meets minimum targets.

2. This model is not going to excite people about the future.

3. This model has no excess capacity for other organisms.

4. This model will end up with only 2-3 funded centers.

5. This model still requires more 1997 $ than we are currently planning on!



NOTES:

1. This model could do the whole genome if necessary! Again excess capacity, once proven, could be used for other organisms.

2. The plateau after year 6 and decrease in $ thereafter will cause folks to leave the game as the "end" approaches.

3. This rapid ramp in capacity may not be feasible and is certainly less cost effective.

4. This rapid ramp in $/year will kill all other investments by 1999.

Draft

NIGMS WORKSHOP

NEW APPROACHES TO THE STUDY OF COMPLEX BIOLOGICAL PROCESSES

November 24-25, 1997

In September 1997, members of the National Advisory General Medical Sciences (NAGMS) Council were informed that staff intended to convene an informal workshop on analyzing complex biological systems. This initiative reflects a growing sense among some researchers that investigators are encountering significant new challenges that may go beyond traditional and even very recently developed biomedical research approaches. Thus, to continue making progress, investigators may well need fundamentally new strategies, approaches, and tools to identify and understand the design principles and dynamics of complex biological processes.

The workshop was convened on November 24-25, 1997, at the National Institutes of Health in Bethesda, Md. Participants included researchers with specialties such as genetics, biochemistry and physiology, the neurosciences, and medicine. Some participants were trained in non-biological disciplines such as mathematics, physics and engineering, with experience in the analysis of complex systems. One workshop participant, Dr. Susan Henry, who is a NAGMS Council member, was asked to deliver the workshop's recommendations to other members of the Council during its January meeting.

DEFINING THE CHALLENGE

During the past decade, biomedical researchers have been amassing an enormous volume of valuable data across a wide spectrum of the biological world. This information ranges from detailed molecular descriptions of multicomponent protein systems, including important enzyme-substrate and receptor-ligand complexes, to genomic DNA sequence information for more than a dozen microorganisms as well as extensive DNA sequence information for other microorganisms, plants, and animals. On another level, biologists are also learning a great deal about essential subcellular structures, such as the mitotic apparatus for separating chromosomes and organelles that are responsible for cellular locomotion, and about the way genetically specified programs operate during differentiation and development of specialized cells, tissues, and organs.

Nearly all these efforts reflect a reductive, analytic approach to investigating important biological questions. Typically this approach entails careful, often intensive experiment-based scrutiny of a very limited number of components in a biological system, model building and hypothesis development based on those empiric observations, and further experimenting to test elements of those hypotheses.

Reflecting the value in following this approach, biomedical researchers from a range of disciplines typically have deliberately restricted their analyses to well-defined systems with relatively few components. However, more recently, that expressly narrow approach is being complemented by

broader, more comprehensive efforts. In particular, expanded programs to use genomic DNA sequences to identify genes and their regulatory sequences, and predict the structure and even the function of their coded proteins, will provide a phenomenal volume of valuable new data about an organism's entire genetic complement.

Equally important, however, these efforts to analyze genomic DNA sequences as well as other large-scale analytical efforts pose an immense challenge to those trying to understand fully what this information means for biology. Thus, useful though it may be, a complete listing of an organism's genetic and structural components is not adequate to describe or explain, much less predict, the behavior of that organism's many complex and varied functions. Many of these functions may result from stochastic rather than fully programmed interactions of genetically specified products. The behavior of the whole may not be inferable from the collective description of individual parts. Much is to be learned regarding how seemingly unrelated molecular events can influence the development of a complex phenotype.

Such realizations prompt a series of challenging questions for biologists to address. Those questions revolve around fundamental issues of how they conduct their scientific investigations and analyze information. For instance, within such comprehensive data sets, which details are essential and what others may safely be disregarded? More important, are there principles to be discovered that will help investigators describe emergent biological properties as they analyze such data sets? If so, how can they begin to identify and then effectively deploy those principles?

Other broad questions surfaced in discussion. For example, can investigative teams, with members drawn from different disciplines, begin to develop "hybrid" approaches to studying complex problems--perhaps by combining traditional bottom-up analysis with reverse engineering strategies? Is a new "integrated" and "reiterative" rather than strictly reductive approach now needed for studying biological systems?

Such questions suggest some special pragmatic needs, and a new initiative. In addition to supporting the development of interdisciplinary research projects, perhaps the most important need will be to develop investigators who can deal with the inherently multidisciplinary nature of such research. Meeting these training needs may well be as challenging culturally as it will be intellectually. In addition, the need for specialized instruments is anticipated as is the need for computer software systems that are capable of integrating large volumes of seemingly unrelated data.

EXAMPLES HELP DELINEATE FRAMEWORK FOR INITIATIVE

Despite the risk that descriptions of specific biological examples might limit the scope of the anticipated NIGMS initiative, workshop participants found these descriptions helpful for defining the initiative's framework. These examples, which are drawn from research on both prokaryotes and eukaryotes, thus provide a concrete sense of what investigators mean when they refer to complexity in biological systems, but they are not meant to constrain the boundaries of the anticipated initiative.

Consider enteric bacteria such as *Escherichia coli* and *Salmonella typhimurium*, both of which have been intensively studied for several decades. Indeed, a great deal of detailed information is available to describe their respective biochemistries, genetics, and physiologies. For example, in 1997, the *E. coli* genomic sequence was completed and published. Nonetheless, great gaps remain in understanding the behavior of these bacteria, with some of those gaps reflecting phenomena that seem to reach beyond a common-sense understanding of their genetic or physiologic functions.

For instance, fully one percent of the *S. typhimurium* genome is dedicated to genes specifying proteins needed to synthesize vitamin B12. Yet, if those genes are deleted, the mutant cells exhibit no obvious phenotype when grown in culture. Do these genes specify some other function that is needed when *S. typhimurium* is growing in a more natural setting? And why does this bacterium carry these genes when its close relative *E. coli* does not? These questions lead to a more fundamental question: What evolutionary strategy underlies the features that distinguish one closely related bacterial species from another?

Microorganisms offer many other examples of biological complexity. Despite decades of intensive study, investigators are far from understanding the transition in *Bacillus subtilis* from vegetative growth to spore formation. Because this transition seems to involve cellular responses to environmental signals, and not all cells within a seemingly uniform population go through it, knowledge of the bacterium's genomic sequence is unlikely to provide an explanation for how this transition process is initiated. Some broader overview of the regulatory circuitry at work in such cells seems a necessary prerequisite for understanding this and other similarly complex biological processes.

Investigators studying bacterial chemotaxis also are faced with the challenge of understanding how living cells transduce and respond to environmental signals. Although a great deal is known about the genetically determined biochemical apparatus that enables bacterial cells to move up or down a chemical gradient, much is yet to be discovered about how the regulatory process functions to produce an appropriate response to information in the environment.

Despite the availability of the genomic sequence of the yeast, Saccharomyces cerevisiae, that comprehensive DNA sequence information seems not to explain several genetic and metabolic phenomena peculiar to living yeast cells. For instance, yeast lipid metabolism follows distinctive patterns during different phases of cell growth. However, although the general category of end-lipid product seems to be under genetic control, the overall process is also affected by other more subtle factors, including catabolite repression, conformational changes of proteins, and the physical state of the plasma membrane. How is this information integrated and processed to determine the outcomes of lipid metabolism?

Another element of biological complexity is what some researchers are calling "not-strong" genetic effects--a term that seems to apply to phenomena associated with mating cell signal transduction in yeast. This process involves a complex cascade of biochemical changes among small signal molecules and kinase proteins, whose overall control may reflect subtle interactions between cell types. Because mutant selection methods are typically biased toward components that have strong biological effects, these other more subtle effects usually are overlooked and

remain difficult to analyze. However, particularly in the context of interacting network systems within cells, these putative not-strong effects may be essential for fine-tuning those systems.

How can investigators identify and study such phenomena? One promising experimental approach entails producing large arrays of microbial cell colonies, each containing a different mutation as well as a fluorescent marker, and then subjecting those arrays systematically to different physical and chemical perturbations. However, such experiments generate voluminous data sets that are proving challenging to analyze in themselves.

Nonetheless, several workshop participants independently recommended this general approach - namely, of subjecting some biological phenomenon to exhaustive testing in many different environments and under many different conditions. This approach provides a way of examining "robustness" of the regulation of physiological processes, according to some investigators whose focus is on microorganisms. Moreover, according to others who are working with complex mammalian systems, such as the genes expressed in the embryonic spinal cord, a similar exhaustive approach may furnish insights into multigenic processes during embryonic and early post-natal development.

Differentiation and development certainly are among the biological processes that investigators deem complex and, for now, elusive. Here again, although genetic studies provide essential insights, they apparently do not tell the complete story. For example, a genetically specified structure that is part of the sexual apparatus in the roundworm, Caenorhabditis elegans, gives rise to part of the visual system in the fruit fly, Drosophila melanogaster. In another instance, gene dosage and slight differences among proteins in a multicompnent complex apparently determine whether an individual C. elegans will be female or hermaphroditic. What accounts for these different outcomes?

Many investigators are now producing "knock-out" mutants in organisms ranging from bacteria to mice as a way of studying complex genetically based behaviors. Yet, despite detailed knowledge about the functions of the targeted genes in such knock-out mutants, often the resulting phenotype deviates from the one anticipated. Typically, seemingly redundant genes with overlapping functions help explain what happens, raising another fundamental question. Why is there so much apparent "redundancy" among genes?

Moving to a clinical setting, multi-organ failure provides an important example of a complex, poorly understood biological phenomenon that often proves deadly and, even when it can be successfully countered, is very expensive to treat. In this clinical situation, several vital organs begin to move away from healthy homeostasis near or at the same time, and toward a state of severe dysfunction that brings death. Typically, although each organ system is treated separately to try to reverse its dysfunction, negative synergy often occurs among several organ systems, meaning that even heroic efforts to treat one deteriorating system may not prevent the others from entering a downward spiral. This life-threatening clinical phenomenon poses a difficult challenge for investigators seeking to better understand and treat patients who develop this syndrome.

FACING CHALLENGES PROMISES BOTH BASIC AND PRACTICAL REWARDS

During the past several decades, biomedical investigators have used ever-more sophisticated research tools to identify and analyze the functions of the components that make up living cells. Despite many successes, they often have met with frustration when they have tried to describe phenomena that embody biological complexity--that is, functions that map across several organizational dimensions. For example, although understanding a monogenic disorder may prove to be relatively straightforward, understanding a multigenic disorder that affects several potentially interacting biochemical pathways and physiological processes usually does not.

A major source of this frustration is the absence of a common language and of compatible (and accessible) data systems for much of the analysis that is needed. To be sure, there is a standardized naming convention for enzymes, and DNA sequence data sets are relatively easily manipulated. However, gene and gene product nomenclature tends to be idiosyncratic at best, making it difficult to compare potentially common structures between any but the most closely related organisms. Indeed, analysis of homologous structures within different organs of a single species, such as a particular molecular apparatus used during development and morphogenesis in the fruit fly, has been hampered because of unstandardized nomenclature.

An NIGMS initiative will likely focus on fundamental issues of biological complexity, including questions of complex multigene and gene product interactions, membrane signal transduction and responses to subtle environmental factors, and of differentiation and development in model systems. There are, however, opportunities at higher levels of organization, in the clinical setting. In the case of multiorgan dysfunction, failure, and death, and other areas of NIGMS concern such as anaestesiology, pharmacology, and burn research, quantitative insights into the function of complex systems in humans could help to improve health, save lives, and reduce the costs of medical care.

RECOMMENDATIONS

Summary

The workshop participants recognized that the emerging broad subject area, provisionally termed "Biological Systems Analysis (BSA)," merited strong NIGMS support and encouragement. The participants emphasized that, although there were differing views on approaches to the analysis of complexity in the context of diverse levels of organization in biological systems, a unifying goal could be identified. They suggested several means to achieve this goal.

The goal is to promote analysis of the design principles and dynamic behaviors of complex biological systems, with the expectation that such an understanding will impact the treatment of human disorders and disease. If successful, these design principles and dynamic behaviors will be presented in quantitative formats that readily allow testing by both computer modeling and *in vivo* approaches.

Currently, there are a number of projects, some of which were presented by participants, that merit inclusion in BSA. However, quantitative, integrative treatments of classical molecular biological, genetic, cell biological and biochemical data are relatively novel, with few experienced

investigators. The participants suggested that adoption of these quantitative approaches would require a variety of supporting initiatives. These fall into the broad categories of cross-disciplinary and collaborative research projects, educational efforts, and infrastructure support.

Specific Recommendations

Support of Cross-Disciplinary Research and Collaborative Research Projects

Opportunities exist for the integration of traditional biological approaches with those of physics, mathematics, chemistry, engineering and computer sciences. These opportunities have resulted, in part, from technological advances that are amplifying our ability to acquire massive and comprehensive datasets of unprecedented resolution in time and space. Furthermore, these opportunities can be found across the scope of science that NIGMS supports, from basic studies with model organisms to clinical studies that impact directly upon the management of human disease. Comparative studies across diverse systems may yield unifying patterns of biological organization and dynamics. The participants therefore encourage NIGMS to announce a program of support for cross-disciplinary and collaborative research projects that will enable us to understand, represent, and predict the behavior of complex biological phenomena.

Support of Educational Efforts

The participants recognized that there are significant educational barriers to the realization of these new research opportunities. One classic barrier is the lack of communication between theorists and their more empirical colleagues within the biomedical community; another is the traditional structure of academic departments that may not be supportive of cross-disciplinary appointments and formal partnerships. In order for the value of BSA to be appreciated within the traditional biomedical research community, the participants recommend that NIGMS support further workshops and scientific meetings that will publicize the promise and accomplishments of BSA.

A major barrier is the shortage of biomedical scientists who also have the quantitative and computational expertise that must be brought to bear on these research areas. The workshop participants recommended several remedies to address this shortage.

First, physicists, mathematicians, engineers, computer scientists, and other experts with quantitative skills relevant to the analysis of complex processes and complex genetic traits should be encouraged to collaborate with biomedical scientists. One example is the NIGMS initiative to provide supplements to the Institute's grants that will allow scientists with these backgrounds to conduct collaborative research projects intended both to develop new approaches to the study of complex systems as well as provide experience for non-biologists in working with biological systems.

NIGMS is encouraged to solicit applications for institutional and individual fellowship applications that will provide relevant cross-disciplinary instruction and research experiences at the pre-and postdoctoral levels. The current NIGMS program, Systems and Integrative Biology, would be appropriate for institutional training grants at the predoctoral level. Postdoctoral

training at both junior and senior levels that emphasizes cross-disciplinary experiences should likewise be encouraged. While the existing Supplements Program, referenced above, can provide some support, the NIH Independent Research Service Awards (fellowships) are an appropriate mechanism. All avenues should be encouraged, perhaps through a Program Announcement.

Support of Resources and Infrastructure

Technological advances and access to data have been key to opening up opportunities for BSA. However, there are significant barriers to acquiring access to expensive instrumentation and to the development of new software and critical databases. Also, access to, and maintenance of, existing databases currently can be a problem. NIGMS should anticipate that these needs will continue to accelerate as more high-throughput, systems-wide approaches are invented and are accepted as routine research tools. In order for the biomedical community as a whole to share in these developments, sources of funding and creative approaches to access will be required.

The workshop participants further propose that the NAGMS Council approve the appended resolution, to be presented to the Director of the NIH, that urges cooperative efforts to provide access to commercial databases of value to publicly funded biomedical research.

Draft

The Genetic Architecture of Complex Traits Report and Recommendations

This report summarizes the findings of a panel of experts who met at the National Institutes of Health on December 10-11, 1997 at a workshop entitled "The Genetic Architecture of Complex Traits." The report and recommendations were prepared for consideration by the National Advisory General Medical Sciences Council.

Executive Summary

Most genetic traits of interest in populations of humans and other organisms are determined by many factors, including genetic and environmental components, which interact in often unpredictable ways. For such complex traits, the whole is not only greater than the sum of its parts, it may be different from the sum of its parts. Thus, complex traits have a genetic architecture that consists of all of the genetic and environmental factors that contribute to the trait, as well as their magnitude and their interactions.

The following recommendations are intended to increase the rate of progress and improve the quality of research on the analysis of complex traits.

Research

Data

- NIH and the scientific community should focus on better acquisition of data, including larger samples and more refined phenotype definitions.
- There is no single simple sampling scheme that should be used to obtain the data for these studies.
- Given the reality that the genetic architecture will and can vary as a function of population parameters, collection of data on population structure, including histories for different human populations and for the model organisms, is essential and should be supported.

Methods

- The richness to the current methods should be exploited to its fullest through creative applications to appropriate data sets.
- Every effort should be made to encourage the development of fundamentally new models and more sensitive methods of analysis.
- Because so little is known about genetic architecture, exploratory and observational studies should be encouraged.
- NIGMS should encourage collaborative studies among investigators in diverse disciplines.
- With respect to review and funding of research grant applications, it is important to emphasize that no one method or model for studying genetic architecture can be adopted universally.

Model Organisms

- The choice of organism and research design should be dictated by the complex trait of interest and the questions being asked.
- Studies of population structure and variability of organisms in natural populations are needed.

Resources

- Support is needed for new database structures for population data.
- The development of publicly available genetic data sets of genetic maps and haplotypes should be encouraged.

Training

- The expertise of computational scientists, including physicists, mathematicians, and engineers, is needed; however, most will need to be retrained in statistical genetics.
- NIH should support training in statistical genetics for scientists who intend to apply the tools of genetic analysis.
- Multidisciplinary training is essential.

Communications

- The analysis of complex traits does not lend itself to quick and easy solutions.
- At the same time, it is vitally important to communicate the results of studies in accurate terminology.

Introduction

Most traits that vary in populations of humans and other organisms are determined by multiple factors. Most common diseases with a genetic component are such complex traits. The complexity arises from the fact that each factor contributes, at most, a modest amount to the total variation in the trait observed in the entire population. Complex traits may be continuous in distribution, like height or blood pressure, or they may be dichotomous, like well and affected. Multiple genetic and environmental factors may interact with each other in unpredictable ways. Such unpredictable, nonlinear interactions mean that the expression of the trait may not be anticipated from knowledge of the individual effects of each of the component factors considered alone, no matter how well understood the separate components may be. Thus, the whole is not only greater than the sum of its parts, the whole may be different from the sum of its parts.

The genetic architecture of complex traits consists of a description of all of the genetic and environmental factors that affect the trait, along with the magnitude of their individual effects and the magnitudes of interactions among the factors. It is, in principle, possible to define the genetic components in terms of Mendelian segregation and location along a genetic map. Environmental factors are much less easily partitioned into separate factors whose individual effects and interactions can be sorted out.

It is critical to recognize that the genetic architecture is less a fundamental biological property of the trait than a characteristic of a trait in a particular population. The genetic architecture is a moving target that changes according to gene and genotype frequencies, the distributions of environmental factors, and such biological properties as age and gender. The dependence on gene frequencies creates some seeming paradoxes of genetic architecture. For example, suppose a trait is completely determined by the interaction of two recessive alleles, one of which is rare and the other common. At the population level, the trait appears to be determined by the rare allele, because its presence limits the variation in the occurrence of the trait among individuals. If the allele frequencies were reversed, the other gene would appear to be the determining genetic cause. But in either instance, both recessive alleles contribute equally to the biological causation of the trait. The implication of the population dependence is that the predominant genetic factors contributing to a complex trait may seem to differ from population to population. This is probably one reason for the apparent heterogeneity sometimes found in the results of genetic linkage studies in different populations. Insufficient statistical power in the linkage tests is also a possible explanation, and there is always the possibility that superficially identical complex traits in different populations may actually have different biological causes.

The existence of unpredictable, nonlinear interactions between the multiple factors affecting complex traits, as well as possible frequency-dependent differences in genetic architecture from one population to the next, emphasizes one of the principal conclusions of the December 10-11 workshop, "The Genetic Architecture of Complex Traits." The

participants unanimously agreed that understanding the genetic and environmental basis of complex traits is not going to be easy and will not be achieved in a foreseeable time frame. Too little is known about the true nature of the complexity of such traits in any organism. In an ideal case, when the factors are not numerous, when their main effects are quite large and their interaction effects quite small, and when interpopulation heterogeneity is minimal, very rapid progress can be made. It is by no means clear how widely actual complex traits in humans and other organisms depart from these ideal conditions. Furthermore, while improved technology can be of tremendous importance, the challenges are not only technological. They are also conceptual (for example, how to identify nonlinear interactions, how to optimize computational algorithms), clinical (improved diagnostic criteria), and epidemiological (how to sample in such a way as to minimize spurious associations due to population structure and population history while maximizing the power to detect biologically significant associations).

Because the genetic architecture is a characteristic of a trait in a population, it is affected by population structure and population history – a fact that undermines the concept of a disease gene. In a complex trait, there is no disease gene in the sense of a Mendelian factor that, by itself, causes a disease. Rather, the genetic and environmental factors underlying a complex trait must each be considered as contributory or predisposing rather than as determinative. Where diseases are concerned, genetic components may be regarded as risk factors.

In spite of these difficulties, the analysis of complex traits is fundamentally important to identifying the contributing genetic and environmental factors of traits and to understanding their underlying biology. The discussion and recommendations from "The Genetic Architecture of Complex Traits" workshop focused on opportunities for progress in four areas – research, resources, training, and communications – some of which can potentially be addressed by NIGMS. Other points will require discussion and action by other groups.

Research

Data

The number of individuals that can be studied will ultimately determine the limit of resolution of analyses of complex traits. The sample sizes that researchers can collect and the quality of individual phenotype assignments (the ability to recognize and correctly assign trait values to individuals) are serious barriers to progress. Given the on-going efforts to produce a dense, quality map of the human genome of single nucleotide polymorphisms (SNPs), sufficient numbers of markers to analyze the complex traits will be available soon. In human studies, therefore, the limitation will be correctly phenotyped individuals – that is, our ability to correctly diagnose disorders or completely describe traits. The situation is not quite the same for various model organisms where the development of new markers and new maps has lagged behind the effort for humans.

- **NIH and the scientific community should focus on better acquisition of data, including larger samples and more refined phenotype definitions.** This is

especially true for studies of specific common diseases for which large samples of individuals who are assessed for disease according to the same criteria are very difficult to acquire.

- **There is no single simple sampling scheme that should be used to obtain the data for these studies.** Data rich does not necessarily mean information rich. The questions being asked should dictate the analysis of choice, the types of data necessary, sampling design, and the final sample size.

- **Given the reality that the genetic architecture can and will vary as a function of population parameters, collection of data on population structure, including histories for different human populations and for the model organisms, is essential.** These data must include information defining normal variation within these populations as well as recording of disease phenotypes. Accumulation of these population data sets will also lead to studies of evolution within the human population and answers to questions about how these traits came to be and how some diseases achieve sufficient population frequency to be common within our population. Because the genetic architecture differs according to how closely the trait is related to survival and reproductive fitness, evolutionary forces can profoundly affect the genetic architecture of complex traits in humans and other organisms.

Methods:

The consensus of the participants at the workshop was that both current methods and fundamentally new approaches should be pursued aggressively.

- **The richness to the current methods should be exploited to its fullest through creative applications to appropriate data sets.** The current armamentarium of methods provides numerous ways to model traits, analyze data, and evaluate results for internal consistency and biological reality. However, just because these techniques and methods will allow us to learn a great deal, we should not hesitate to expand and develop them, to refine the models, and to improve them especially with respect to computational limitations.

- **Every effort should be made to encourage the development of fundamentally new models and more sensitive methods of analysis.** Human geneticists have been slow to explore and adopt methods developed in other areas of science. For example, refinement of techniques for more rapid computation should be actively pursued. NIH, in its interaction with theoretical investigators and its review process, should encourage creative efforts even if they are not guaranteed to provide improvement.

- **Because so little is known about genetic architecture, exploratory and observational studies should be encouraged.** Although the dogma at NIH is that only hypothesis driven research has merit, science is, in fact, built on observations. We are sufficiently naïve about complex traits that exploratory studies must be supported.

- **NIGMS should encourage collaborative studies among investigators in diverse disciplines.** Complex traits are not the province of any single discipline. The expertise of molecular biologists, biochemists, clinicians, evolutionary biologists, developmental biologists, mathematical and statistical geneticists, and many others is needed.

- **With respect to review and funding of research grant applications, it is important to emphasize that no one method or model for studying genetic architecture can be adopted universally.** No one method can answer all, or even some, of the questions without being using in concert with additional approaches. It is a serious error to insist that all studies apply one method, such as association studies for linkage, when the full suite of methods is available to the investigators. In addition, it is inappropriate to abandon the candidate gene approach in favor of general genome searches for those traits where reasonable biological or physiological candidates can be identified. It is also inappropriate to insist on a candidate gene approach when a whole-genome search is justified by sample size and cost-effectiveness. Significant information has accrued both through the exploration of candidate gene regions and the rejection of candidate gene hypotheses.

Model Organisms

The overriding advantage for model organisms is the ability to do both genetic and environmental manipulation that can not be done with human beings. Studies using animal models to explore the genetic architecture of complex traits should be supported in order to identify general principles and pathways and to gain broad understanding of the biology of complex traits.

- **The choice of organism and research design should be dictated by the complex trait of interest and the questions being asked.** There is no single, limited set of organisms that is sufficient for these studies. Studies of non-traditional organisms may have much to contribute.

- **Studies of population structure and variability of organisms in natural populations are needed.** Because the genetic architecture of complex traits varies with context, measuring traits and identifying their variability in the wild is an important piece of the puzzle, and studies to do so have almost ceased to exist.

Resources

- **Support is needed for new database structures for population data.** A great deal of population data already exists, but most public databases, such as Genbank, are inadequate for handling population data, which must include the population source of the allele sequences and the frequency with which they occur. Many data currently exist on individual investigators' computers, but it is not accessible to other scientists because there are no good mechanisms for sharing data. The study of population structure would be greatly enhanced by establishing one or more databases to make

these data easily and reliably accessible. Pilot efforts might begin by augmenting GDB for human data or FlyBase for Drosophila data.

- **The development of publicly available genetic data sets of genetic maps and haplotypes should be encouraged.** There are numerous NIH-supported studies that individually have low power to detect and map factors that contribute to complex traits. NIGMS should support the establishment of databases that enable data from such studies to be combined. Further, many successful gene mapping studies generate marker data that could be useful to other investigators and that could be made publicly available once the gene of interest is mapped.

Training

- **The expertise of computational scientists, including physicists, mathematicians, and engineers, is needed; however, most will need to be retrained in statistical genetics.** One of the impediments to such recruitment is the relatively low salary scale of entry-level postdoctoral students in biology.

- **NIH should support training in statistical genetics for scientists who intend to apply the tools of genetic analysis.** Statistical methodologies must be applied knowledgeably, especially where human data are concerned. Many scientists, including clinicians, molecular geneticists, and others, are eager for basic training in analytical methods so that they can collaborate effectively with their statistical colleagues.

- **Multidisciplinary training is essential.** Studies of complex traits are inherently multidisciplinary, requiring expertise in genetics, statistics, computational biology, and other areas. In human studies, a strong clinical component is often essential because of the need for careful and correct diagnoses. All too often there are structural or institutional barriers to multidisciplinary collaboration and training; nevertheless, students must be prepared to cross disciplinary boundaries if they are to succeed and contribute to future research studies.

Communications

The genes involved in complex traits are contributing factors rather than disease genes. Any one of the genetic factors that contribute to a complex trait may actually account for a relatively small proportion of the total variation in the trait. Furthermore, by itself, the gene may not cause the disease, but rather may be one of many contributing genetic and environmental factors. The danger of oversimplification is to mislead the public into thinking that a disease has been conquered and effective new treatments and therapies are just around the corner. There is a great danger in publicly raising false hopes.

- **The analysis of complex traits does not lend itself to quick and easy solutions.** We do not yet know the true degree of complexity of complex traits. Hopefully, some will approximate the simplest ideal case and be analyzed rather quickly, but others

will be much more difficult. It is prudent to make no promises until we understand the nature and extent of complexity in genetic systems.

- **At the same time, it is vitally important to communicate the results of studies in accurate terminology.** There is no gene for hypertension, depressive disorder, obesity, or any other complex trait. All genes that affect the trait are contributing factors that are more or less important only in relation to other contributing genetic and environmental factors in defined populations. Likewise, terms such as "heritability" are misleading since, in normal language, the term implies transmissibility not the technical meaning of a ratio of variances. The public as well as our scientific and medical colleagues needs current and accurate information in order to understand the issues surrounding the study of complex traits. Scientists should take great care to communicate clearly in order to promote understanding of these difficult and very important issues.

Conclusions

The findings of the participants at "The Genetic Architecture of Complex Traits" workshop do not lend themselves to simplistic answers or quick fixes. The recommendations need to be considered thoughtfully and thoroughly, sometimes in collaboration with a broad spectrum of the scientific community. Success will depend on coordination among institutes and agencies and on increased understanding of the complexities of the scientific questions being asked. The participants note that NIGMS can address many issues; however, some recommendations are trans-NIH. The National Advisory General Medical Sciences Council may wish to consider broader dissemination of the report and recommendations.

A Resource for Discovering Human DNA Polymorphisms

The US is a nation of immigrants. We are dealing with ethnicity by considering the geographic region of the ancestors of the individuals. Admixture among the groups means that many minority Americans have some European ancestry. NHGRI is still consulting with anthropologists about the amount of admixture in various groups, but the first estimates are given in the second column below. This table shows the numbers of individuals from the groups that will be sampled, and what proportion of their ancestry comes from the four major geographic regions of the world. In particular, European ancestry is included in all the other groups, so samples of African-Americans, for example, have about 85% African ancestry and 15% European ancestry. The numbers of individuals in the various groups were chosen to be roughly equal. Compared with Europeans and Asians, Africans have more variation and Native Americans have less, so those numbers were adjusted accordingly. Also, since the US is predominantly of European ancestry, the number of individuals of European ancestry was increased. We are still considering the exact numbers, but they are roughly:

	Proportion European	Africa	America	Asia	Europe	Number
European-American	1.00	0	0	0	57	57
African-American	0.15	125	0	0	22	147
Mexican-American	0.75	0	5	0	15	20
Hawaii	0.10	0	0	62	7	69
India-American	0.10	0	0	60	7	67
Alaskan Native	0.10	0	38	0	4	42
Southwest Amerindian	0.25	0	37	0	12	49
Mixed Central Amerindian	0.25	0	37	0	12	49
		125	117	122	136	500