

Recommendations for Future Efforts in Primate Genomics

I. Introduction

In January 2001, a group of researchers who use nonhuman primate models of human diseases, National Institutes of Health (NIH) staff, and regional primate center directors and staff met in Seattle, WA to discuss current scientific information and share their knowledge regarding the genomics of nonhuman primates. The goals of this workshop, which was organized by the Washington Regional Primate Research Center and National Center for Research Resources, were a) to assess the information presently available concerning primate genomics, b) to consider what information and resources would be most valuable in facilitating future research progress in this area, and c) to discuss how the eight regional primate research centers should contribute to future research initiatives in primate genetics and genomics.

The attendees included Dr. Judith Vaitukaitis, Director of the National Center for Research Resources (NCRR); Dr. Jerry Robinson, NCRR Program Director for the Regional Primate Centers; staff from the National Human Genome Research Institute (Dr. Mark Guyer) and Office of AIDS Research (Dr. Marta Leon-Monzon); the Directors of the eight regional primate centers; other primate center staff; representatives from the human genomics research community and the commercial biotechnology industry; and other interested experts in primatology, genomics, and animal models of human diseases.

There was broad and strong consensus that extensive analysis of genome structure and function in selected nonhuman primates could make immediate and significant contributions to the overall mission of NIH by accelerating progress in the understanding of many human diseases. A nearly complete sequence of the human genome is now available, and genomics is playing an increasingly important role in biomedical research. Studies of primate genomics will become an important parallel and adjunct to human genomic research. Primate models of human diseases, such as atherosclerosis, AIDS, diabetes, osteoporosis, neurodegeneration, mental illness, alcohol dependency, asthma, cancer, and others are critical to the long-term success of biomedical research.

The availability of genetic information will enhance the value of nonhuman primates, both as models for specific disease processes and as tools for understanding normal physiological processes. Rhesus monkeys, baboons, chimpanzees and other primates are also used to test new therapies and to answer fundamental questions in mammalian biology that cannot be addressed directly in humans, or effectively in smaller species such as rodents.

This report summarizes the discussions and recommendations of the group that met in Seattle. We strongly believe that there is an immediate need to expanded efforts to understand the genome structure and function in several primate species. Major effort should be directed toward generating the complete DNA sequence of the rhesus macaque (*Macaca mulatta*), extensive transcription mapping in rhesus monkeys, baboons (*Papio hamadryas*) and chimpanzees (*Pan troglodytes*), and additional physical and genetic linkage mapping for baboons and rhesus macaques. It is also important that funding be available for functional genomic studies such as analyzing differential gene expression, using expression array technology, and investigating gene function through proteomics. The large amount of data to be generated will require appropriate bioinformatics resources and personnel to ensure that the data can be accessed and used efficiently. Our specific recommendations for priority funding (See Table One.) follow a general discussion of the rationale

for expanding support for this line of research.

Table One: Summary of Major Recommendations and Goals

Basic Resources Needed:

- BAC libraries for rhesus monkey, baboon, and chimpanzee
- cDNA libraries from 8-10 tissue types or developmental stages for rhesus monkey, baboon, and chimpanzee

Genome Mapping:

- Genetic linkage maps with 3 cM density for rhesus monkey and baboon
- Radiation hybrid maps of at least 5,000 markers for rhesus monkey and baboon
- BAC contigs generated by BAC end sequencing for rhesus, baboon and chimpanzee

Sequencing:

- cDNA sequences from 8-10 tissue types or developmental stages each from rhesus, baboon and chimpanzee
- Complete shotgun sequencing of the rhesus monkey genome

Bioinformatics:

- Databases and required personnel for information services within eight regional primate research centers to support research
- Specialized databases for genome informatics and personnel to annotate the information on nonhuman primate DNA sequences, genome maps, expression array data and other resulting information

Functional Genomics:

Expression Array Analyses

- Support for analyses of gene expression in nonhuman primates using human cDNA arrays
- Support for development of species specific expression arrays where needed
- Support for exploration of novel applications of expression array methods in nonhuman primates

Proteomics

- Support for the creation or expansion of two or more proteomics centers that will provide protein isolation, identification, and quantification support to research in the nonhuman primates.

II. How has the Human Genome Project benefited biomedical research?

The sequencing and functional analysis of the human genome is transforming biological research. Our understanding of human biology and disease is undergoing a rapid and fundamental revolution as a result of the determination of the human DNA sequence. Future advances in biomedical research concerning the causes and treatment of human disease will be accelerated through continuing elaboration and interpretation of this information.

There are several reasons why human genomic research is having an extraordinary impact on efforts to understand and treat disease. First, many disorders--including Duchenne's muscular dystrophy, Huntington's disease, fragile X syndrome and many disorders of metabolism--are entirely or essentially genetic diseases. These disorders can be understood only when the molecular genetic basis of pathology is discovered. Often, the causative mutation occurs in a single gene, and may be as simple as a single altered nucleotide. However,

an increasing number of *single-gene* disorders are shown to result from large genomic insertions, deletions or rearrangements involving multiple genes, rather than simple point mutations of nucleotides. Data on genome organization and structure, as well as mechanisms of mutation, provide the essential groundwork for elucidating the causes of Prader-Willi syndrome, Smith-Magenis syndrome, and a continuously expanding list of other disorders.

Delineating the causes of other diseases (e.g., cancer, osteoporosis, hypertension, substance abuse, et al) requires us to address both inherited genetic susceptibility and external factors such as exposure to environmental insults, infection by pathogens, or personal choices about diet and lifestyle. Interaction among genes and environmental factors such as diet, smoking, or exercise are fundamental to many disease processes, and these interactions must be understood at the most basic molecular and cellular levels. For some of the most common human diseases (e.g. coronary heart disease and obesity), the clinical problem is best conceived as an extreme instance drawn from an underlying broad range of normal physiological variation. We cannot understand susceptibility to these common complex diseases until the genetic basis of both normal variation and pathological extremes is explained.

There are other aspects of the biomedical sciences that also benefit from genomics information. The new field of pharmacogenetics is focused on describing the genetic basis of differential response to drug therapies. It is well known that different people with the same clinical condition can respond differently to the same treatment. This variation in efficacy--or adverse effects--is due to genetic differences in drug metabolism. Progress in pharmacogenetics is accelerating rapidly due to increasingly detailed information about the structure and function of genes whose protein products influence these processes. Gene therapy is another aspect of medical research that depends on knowledge of genetics and genomics. Development of improved gene therapy strategies and reagents is made possible through this research.

Finally, recent developments in technology allow researchers to characterize mRNA expression levels, protein expression levels, and protein-protein interactions on a genome-wide scale. The availability of sequence information and/or cDNA clones allows the construction of DNA arrays that can be used to simultaneously monitor the quantitative levels of expression of thousands of functional genes. This approach is being used to examine the different patterns of gene expression between healthy and diseased tissue in remarkable detail. We expect that expression array methods will be used in a variety of ways to investigate the molecular basis of many diseases. The availability of sequence data enables mass spectrometric analysis of proteins and allows similar types of expression studies to be carried out in the protein domain. Genomic sequence data coupled with mass spectrometric proteomics technology enables the identification and characterization of protein post-translational modifications. In addition, the availability of genomic sequence data allows one to rapidly clone and express proteins of interest for further functional characterization (e.g. protein-protein interactions, tissue and cellular localization, identification of important domains, etc.). We believe the most important outcomes of a nonhuman primate genome project will be derived through these and other functional genomics technologies.

III. Why should we examine the genomics of nonhuman primates?

Primates are the closest living relatives of humans, both in evolutionary and genetic terms. Chimpanzees (*Pan troglodytes*) are the animals most similar to humans in overall DNA sequence, with a difference between the two species of approximately 1-1.5%. The other apes, including gorillas and orangutans are nearly as similar to humans. The animals next most closely related to humans are the Old World monkeys, superfamily Cercopithecoidea. This evolutionary group includes the common laboratory species of the rhesus macaque (*Macaca mulatta*), baboon (*Papio hamadryas*), pig-tailed macaque (*Macaca nemestrina*), and African green monkey or vervet (*Chlorocebus aethiops*). Squirrel monkeys, tamarins, marmosets, and owl

monkeys are all New World primates (*Platyrrhini*). New World species are also important in biomedical research, but they are more distantly related to humans than are the Old World monkeys (e.g., baboons and macaques). Consequently, there are larger genetic and physiological differences between humans and New World primates than between humans and Old World monkeys and apes.

Due to our close evolutionary relationship to nonhuman primates, we share many basic features of genetics, development, physiology, and metabolism. These various similarities at the levels of whole-body physiology and metabolism, organ function, cell structure, and even gene organization make primates excellent models for studies of human health and disease.

Many specific and biomedically significant commonalities among humans and nonhuman primates could be cited as illustrations. For example, only chimpanzees can be infected with human hepatitis viruses or HIV. However, because chimpanzees do not become ill as a result of these infections, the best animal model for studying pathogenesis and disease progression in AIDS is the rhesus monkey, which can be infected with simian immunodeficiency virus (SIV) and exhibits pathology remarkably similar to human AIDS.

The neurobiology and brain structure of apes and Old World monkeys are much more similar to humans than are those of other mammals. As a result, primates are very useful models for investigation of complex behavior and cognition, and mental illnesses such as anxiety and depression, drug addiction, and other neurological or behavioral disorders.

Only primates, including Old World monkeys and apes, undergo age-related female reproductive decline and menopause as humans do. Consequently, primates can be used to test the effectiveness of new treatments for osteoporosis, menopause-related lipoprotein changes, osteoarthritis and other degenerative diseases of aging.

Documented cases of monkeys and apes that suffer Down's syndrome, asthma, familial hypercholesterolemia, osteoporosis, diabetes, obesity and lymphoma illustrate the similarities between primates and humans that make these animals excellent resources for many important biomedical studies.

Among biomedical researchers, there is great interest in pharmacogenetics. A clear understanding of how genetic variation among individuals influences the effectiveness of any given pharmacological treatment is both critical to finding the best treatment for any given patient, and valuable in interpreting the results of large clinical trials of new drugs. Genetic analysis of drug metabolism in primates will be an effective component of pharmacogenetics, since the genes that influence drug metabolism among people will function similarly or identically in many primates.

The greatest public health problems in the United States involve common complex diseases such as coronary heart disease, diabetes and hypertension. In these disorders, underlying genetic susceptibility interacts with diet, physical activity, and other aspects of lifestyle to control onset or progression of the clinical condition. By combining improved genomic and genetic information about primates with our ability to experimentally control or monitor diet, exercise, and other non-genetic factors, we can investigate how these genotype-environment interactions occur and understand their impact on disease progression. Studies in which both genetic and environmental factors are under the control of the investigator are, of course, impossible in human subjects.

Finally, transgenesis is a tool that is now being applied to nonhuman primates. Transgenic mice have been an extraordinarily powerful tool for examining the function of and interactions among genes. With the recent success in inserting a foreign gene into rhesus monkeys, it may be possible in the near future to exploit this technology in the context of primate biomedical studies. Additional information about primate genomics is likely to accelerate this progress.

IV. How much is already known about nonhuman primate DNA sequences?

While the human genome project and other genome projects related to model organisms are very well developed or even complete, surprisingly little sequence information has been acquired for nonhuman primates. As of January 2001, there were approximately 4.2 million entries of human DNA sequence in GenBank. However, there were only 10,589 sequence entries for **all nonhuman primate species combined**. Thus, the primate data constitute about one-quarter of 1% of the human total. For rhesus macaque, only 1,297 sequence entries were present, and--due to redundancy in the sequences--these represent approximately 260 unique genes. Given the importance of nonhuman primates in biomedical research and the fundamental significance of sequence data, it is critical that this paucity of genome information concerning nonhuman primates be addressed.

V. Which species should be the primary focus of this expanded effort in primate genomics?

Four factors should be considered when deciding which nonhuman primate species should be the focus (or foci) of effort in an expanded program of genomics research.

First, any species to be investigated intensively should be widely used in studies of several human diseases and be readily available to many interested investigators, including researchers both inside and outside the regional primate research centers program.

Second, the species chosen should be available at low enough cost, in sufficient numbers and in appropriate age-sex classes to permit extensive use as an experimental model. One of the major benefits of the program proposed below will be the resulting opportunity to combine detailed genomic data (e.g., sequence data for multiple functional genes) with experimental manipulation of diet, exposure to pathogens, or other environmental factors in order to test the effects of specific genetic mutations on disease susceptibility and/or progression. This cannot be accomplished unless researchers have access to both the appropriate genetic and genomic data, and the appropriate animal subjects for experimentation.

Third, the species should be suitable for development as a model for additional studies not already underway. We expect the genomic data to be produced will allow researchers to expand the range of studies undertaken, and we wish the major primate species to be suitable for such an increasing role in the biomedical enterprise.

Fourth, because developmental and embryological processes are critical to the origin of many diseases, the subjects of intensive genomics research should be suitable for future studies of genetic influences on embryonic and fetal development. Clearly, human genomic knowledge cannot be applied routinely to studies of human embryos and fetuses, but studies of nonhuman primate development can become an important surrogate for such work.

The recommendation of this group is that the primary effort in nonhuman primate genomics should be directed to the study of rhesus macaques (*Macaca mulatta*). Historically, rhesus macaques have been among the most intensively studied of nonhuman primates. In recent years, more than twice as many rhesus monkeys have been available for--or used in--biomedical research than any other single primate species. Rhesus monkeys are available from a number of breeding institutions across the country, and from non-U.S. breeding facilities as well. It is expected that the number of rhesus macaques used in research each year will continue to increase. While availability may lag behind demand, there is an effort underway to expand the number of rhesus monkeys (especially specific pathogen-free monkeys) available for research over the next 5-10 years.

This species has regularly been used for a wide variety of scientific investigations, including research on vaccines for AIDS, neurobiology and mental illness, immunology, reproductive biology, aging, cardiovascular disease, diabetes, and many others. A literature search of the PubMed electronic database using "rhesus" as the search term found 703 citations for 1998, and 745 citations for 1999. By contrast, the database produced 312 citations for "baboon" in 1998; 318 for "baboon" in 1999; and less than 250 citations for chimpanzees in each of the same years.

The national resource of rhesus monkeys includes several large colonies, now consisting of multiple generations. Specific pathogen-free (SPF) rhesus monkeys are already available, and their numbers are likely to increase in the near term. This species has been studied from the perspectives of biochemical and molecular genetics for many years, though little work in the field of genomics has been published. Through NHGI and OAR, a project to develop a BAC library has been funded, and a recent NCRN grant was awarded to support initial efforts to construct a 10 centiMorgan genetic linkage map over the next four years. This effort will develop a strong framework map but will not provide the detailed resolution required for fine-structure mapping and positional cloning.

Given its primary importance, the entire genome of rhesus macaques should be sequenced as soon as possible. A strategy for accomplishing this goal efficiently is presented in detail below. In addition to large-scale sequencing, a 2-3 centiMorgan linkage map based on human microsatellite loci should be developed, and a radiation hybrid map with 100 - 1,000 kb resolution, consisting of at least 5,000 framework loci (ESTs, STRs and SNPs) should be generated. These maps can be developed quickly and efficiently, focusing on regions that will immediately assist researchers in on-going projects. Because genomic sequence for particular chromosomal regions will be completed randomly, focused attention to map construction will be the most expedient assistance to gene mapping projects. Finally, cDNA clones from a range of tissues and developmental stages should be sequenced and mapped to chromosomal regions.

VI. What information and resources are needed beyond sequence and mapping data?

Assuming that a full sequence for at least one nonhuman primate species becomes available in the near future, and that the cDNA sequencing proposed below for several species also makes significant progress, we expect that many new opportunities to exploit the primate data in novel ways will be created. Analyses that utilize data concerning gene and genome structure, but which address more sophisticated questions related to gene and protein function, are variously described as functional genomics or post-genomic studies. The major insights regarding human health and disease will come from extending the primate genomics program into the realm of functional genomics.

One of the most powerful approaches for exploring gene function is expression array technology. It is now possible, using any of several alternative methods, to investigate the expression of thousands of genes in a given tissue or cell type simultaneously. We anticipate that researchers investigating a wide range of disease processes will combine new nonhuman primate genomic data with expression array technologies to learn more about the molecular causes or consequences of disease. Expression array technologies can also be used to monitor animal response to treatment or to identify correlates of disease progression.

It is relatively straightforward to propose a research plan for the analysis of nonhuman primate genome structure. While we expect that the subsequent post-genomic studies that make use of the information will be central to future biomedical advances, it is more difficult at this time to specify the type of post-genomic research that will be most important. Functional genomic studies will generally be driven by specific hypotheses concerning the causes and mechanisms of particular diseases. However, the group assembled in Seattle felt strongly that, along with funding for genome mapping and sequencing, funding should be

allocated immediately to support initial efforts to apply gene expression array methods and other aspects of functional genomic technology to the study of nonhuman primates.

The details of species, tissue types, and genes to be examined cannot be predicted here, although rhesus macaques, baboons, and chimpanzees will certainly be critical. Funding will be needed to support functional genomic research in these species. The initial studies will likely include: a) evaluation of the use of expression arrays consisting of human clones to investigate gene expression patterns in nonhuman primates, b) the development of species-specific expression arrays for macaques and other primates when warranted, c) the development of new strategies for using expression array methods to facilitate or accelerate identification of functional mutations in genetic linkage studies and d) the creation (or expansion) of proteomics facilities to provide for the investigation of protein expression and interactions in nonhuman primates.

VII. Which additional species should receive special attention?

The group assembled in Seattle felt strongly that the baboon (*Papio hamadryas*) is the second most important primate species for genomics research. Baboons are regularly used as animal models for studies of coronary heart disease and atherosclerosis, osteoporosis, and other human diseases. Baboons are also valuable for studies of various infectious diseases and basic immunology. While not as widely used as rhesus macaques, baboons are among the most frequently used primates in biomedical research (See the results of PubMed searches presented above.), and there is opportunity for this usage to increase in the near future.

Another important factor in favor of expanded genomics research using *Papio hamadryas* is that this species has already received substantial attention from geneticists. Baboons are the best-developed primate species in terms of existing genomic data. A substantial number of genes have been sequenced from baboons, the karyotype is well described, and advanced chromosome painting studies have documented the homologies among baboon and human chromosomes. Most importantly, a genetic linkage map including all the baboon autosomes, with average spacing among loci of less than 7 centiMorgans, is now available. This map was constructed by analyzing more than 320 human microsatellite polymorphisms in approximately 700 pedigreed baboons, using standard computer algorithms to perform marker-to-marker linkage analysis. As a result, it is now possible to conduct full genome scans in families of pedigreed baboons, and such studies have already identified specific chromosomal regions that contain genes of biomedical significance.

It was the opinion of the assembled group that this momentum in baboon genomics should be maintained, though it is not necessary to generate the complete sequence of this species in the proposed initial research. A complete BAC contig map, based on analysis of BAC end sequences will be sufficient for rapid progress, assuming that the complete sequence of rhesus is available for comparison. A 2-3 centiMorgan genetic linkage map, a radiation hybrid map of at least 5,000 loci, and a cDNA expression map using cDNAs from 8-10 tissue types or developmental stages should also be produced as soon as possible.

In addition to studies of the two Old World monkeys, increased information about genome structure and function in chimpanzees (*Pan troglodytes*) would be of significant importance to the biomedical enterprise. Chimpanzees are not available in large numbers, and the cost of use is much higher than for macaques or baboons. It is unlikely that chimpanzees will ever be used extensively for studies of developmental genetics or experimental analysis of gene function. However, no living animal is more similar genetically to humans than are chimpanzees. For some types of analyses, only chimpanzees are suitable as an animal model (e.g., testing of vaccines and drugs against human hepatitis viruses or HIV). As a result, there is clear and compelling reason to invest in additional genomic analysis of this ape species. In addition, the close evolutionary relationship between chimpanzees and humans will provide opportunity for identification and interpretation of sequence differences observed among rhesus monkeys, baboons, chimpanzees, and humans.

Investment in the genomic analysis of chimpanzees is clearly justified, but the scale of effort should be smaller than for rhesus monkeys and baboons. We recommend that a complete physical map for *Pan troglodytes* be constructed by using BAC end sequences for contig assembly and alignment to the human physical map. In addition, extensive cDNA libraries from a variety of tissues should be cloned and sequenced to generate a broad and detailed chimpanzee transcription map.

VIII. Informatics for the NHP Genome Project

Genomic analysis is an inherently data-intensive field. Effective distribution and application of the nonhuman primate genomic information produced by the proposed program will require both computer databases and personnel trained to design and manage them. We anticipate that there will be two distinct but interconnected types of informatics support required.

First, computing resources and database operations will be needed within the eight regional primate research centers. Over the years, each of the eight RPRCs has invested substantially in computing and databases in conjunction with management of their animal resources. Traditionally, these efforts have involved management of animal inventories, clinical and experimental histories, breeding records, etc. More recently, these efforts have been expanded to include management and analyses of pedigrees and genotypic data as a consequence of the increased emphasis on genetic research in NHPs.

These operations are not thought of as informatics in the sense that the term has come to be used in a genomics context; however, it is important to acknowledge the following consideration:

- a. The Primate Centers (and their databases) are, and will continue to be, a major repository for phenotypic, environmental, and pedigree data critical for research on the genetic bases of complex disease.
- b. The Primate Centers will need some level of local genomic informatics capability and expertise if they are to interface with any centralized NHP genomic informatics facility and with scientists who will use Primate Center data in their genomic research.
- c. Existing Primate Center computing and database staff will be responsible for incorporating genomic informatics into their operations. For this to be effective, provision must be made for hiring personnel with genomic informatics expertise and for training staff to deal with new data and methods, so that informatics (in the most general sense) will be a single integrated entity at the centers.

Second, there will be a substantial need for informatics that is traditionally associated with large-scale genomics projects. This will include sequence databases, and substantial capacity to store information annotating and interpreting that sequence. In addition, information regarding gene maps (both genetic linkage maps and radiation hybrid maps) and the homology between nonhuman primate genomes and the human genome will be required. Studies of patterns of gene expression, across tissues and throughout growth and development, will also require specialized databases to provide archive and distribution capabilities. Eventually, proteomics data should be integrated with the more quickly generated data described above.

IX. The proposed research plan will produce several secondary benefits.

The primary motivation for this program in nonhuman primate genomics is to advance the study of human

health and disease. In addition to this primary benefit, we anticipate that the information to be generated will also have secondary applications. One major outcome will be a dramatic increase in our understanding of the basic biology of Old World monkeys and apes. This will most likely lead to improved understanding of diseases and maladies suffered by these animals in captive colonies. We expect that the more detailed knowledge of primate biology will lead to advances in the husbandry, captive care, and management of our critical primate resources. In addition, the genetic information to be obtained will provide much more detailed understanding of both the similarities among humans and our nonhuman relatives and the differences between us. It may be possible to generate hypothesis concerning the genetic basis of unique human traits by comparison with our chimpanzee relatives. Finally, it is clear that many primates are threatened or endangered. We expect that the tremendous increase in knowledge concerning primate biology and genetics may have a positive impact on efforts to conserve these remarkable species.

Proposed Research Plan

A. Overview

The research plan outlined below is based primarily on a proposal presented at the workshop by Leroy Hood of the Institute for Systems Biology. This plan was expanded as a result of discussions among participants. The Executive Committee recognizes that if a nonhuman primate (NHP) genome project is funded, rapidly changing methodologies, techniques, and technologies may suggest alternate approaches that are more efficient or more valuable at the time of funding. Hence, the proposed research plan should be viewed as a working document that provides a broad outline and initial recommendations for a NHP genome project.

The research plan has as its primary goals:

1. The generation of resources for subsequent targeted and whole genome sequencing in the primate species that are most important for biomedical research.
2. The generation of cDNA sequences for selected primate species in order to provide information on the tissue-specific and developmental-stage-specific expression of functional genes.
3. The development of detailed radiation hybrid maps to facilitate construction of NHP physical maps and comparison with human physical maps.
4. Further development of NHP genetic linkage maps to support future efforts in genetic linkage mapping and analysis of primate disease models.
5. The generation of a complete genome sequence for rhesus macaques.
6. Development of reagents and resources to support further studies of gene expression in nonhuman primates using various expression array technologies.

B. General Considerations for Sequencing

The completion of the Human Genome Project, coupled with high degrees of DNA sequence and

chromosomal similarity between humans and nonhuman primates, enormously simplifies the effort required for a NHP genome project.

First, we anticipate that most DNA sequences obtained from a given species (either chimpanzees, rhesus monkeys, or baboons) could be mapped relative to each other by simply comparing the NHP sequences to the human genome. Hence, the assembly of sequence data obtained from a shotgun sequencing approach (the sequencing of random clones) is simplified by the availability of a complete human sequence.

Second, the Sequence Tagged Connector (STC) approach of end-sequencing large insert clones (e.g., bacterial artificial chromosomes or BACs) allows one to rapidly produce a physical map for these NHP genomes by placing the STC sequences on the human map. Chromosomal regions that are duplicated or rearranged in one or more of the NHP species may require additional effort, and assembly will benefit from radiation hybrid mapping data, but the majority of BAC clones will readily be located by comparison to human sequence.

Third, more than 300 genetic markers (primarily microsatellites or simple sequence repeats) have already been mapped in baboons, and a similar effort is underway in rhesus macaques.

Finally, cDNA sequences obtained from a NHP can be mapped directly onto the human genome sequence by searching the human databases for sequence matches. Overall, the availability of the human sequence will allow a NHP genome project to proceed more rapidly and more efficiently than would otherwise be possible.

C. What type(s) and how much sequencing to do

Several different approaches to obtaining sequence data have been used in the human and other genome projects. The human genome project has consisted of a mix of cDNA/EST sequencing, physical and genetic mapping, targeted genomic and chromosomal sequencing, and finally large-scale total genome shotgun sequencing as implemented by Celera. A key consideration for a NHP genome project is the selection, prioritization, and mixing of sequencing strategies and approaches. A NHP genome project must trade off between the degree of completion (i.e., percent of the genome and/or cDNA's sequenced) and cost. It is also reasonable to assume that at some point, the complete sequence of several--or perhaps many--nonhuman primate genomes will be obtained. Our proposal attempts to balance the production of resources that will facilitate subsequent genomics research in biomedically significant species with the immediate need for significant amounts of sequence data on one primary species. In addition, we propose that significant resources should be allocated to support related expression array analysis, bioinformatics and post-genomics efforts that build on an initial mapping and sequencing effort.

As a first step, we propose that funding should be made available to allow the creation of 20 fold redundancy BAC (Bacterial Artificial Chromosome) libraries for rhesus macaque, baboon, and chimpanzee (thus supporting the current efforts of NHGI and OAR). BACs are generally acknowledged to be the premier large insert cloning system for genomics research. The creation of high redundancy BAC libraries for these three NHP species is an essential requirement for subsequent genomics research. These libraries should be arrayed on nylon membranes (filters) and both the filters and the libraries should be readily available to researchers. This will allow researchers to rapidly screen for large insert clones containing genes or markers of interest.

Second, we propose that resources should be made available to end sequence approximately 300,000 BACs from each of the three libraries (rhesus macaque, baboon, and chimpanzee). BAC end sequencing combined with the human sequence information (and knowledge of chromosomal homologies among humans and NHPs) will allow the rapid construction of physical maps for all three species. The BAC end-sequencing approach has been successfully applied to the genomic sequencing of the rice and maize genomes, and it is a

key resource for the assembly of the human genome data produced by Celera. The availability of physical maps generated from BAC end-sequencing will permit efficient targeted sequencing of biomedically significant regions in nonhuman primates and the generation of improved comparative maps.

Third, we propose extensive cDNA/EST sequencing of cDNA libraries generated from selected tissues of nonhuman primates. cDNA sequencing is the most cost effective way to generate sequence information from coding sequences; it also provides information about tissue-specific expression of identifiable genes. We note that nonhuman primate cDNA and EST analyses provide a unique opportunity to construct and sequence libraries from fetal tissues in various developmental stages and perhaps obtain sequences for expressed genes that have not yet been obtained in humans. We feel it is critical to sequence cDNAs/ESTs from at least three nonhuman primates (rhesus macaques, baboons, and chimpanzees) to better interpret sequence differences between humans and nonhuman primates. Full-length cDNA sequences should be obtained as often as is practical.

To minimize the degree of redundant sequencing, we suggest that a process of negative cDNA selection be applied. Briefly, one begins by selecting a small number of cDNA clones from the library/libraries of interest and sequencing them. Sequence analysis is used to identify the unique clones. PCR products are prepared from a representative of each unique sequence and the resulting PCR products are pooled, labeled, and hybridized to large filter-based arrays containing the cDNA libraries. Clones on the filters that do not hybridize to the probe are selected for subsequent sequencing, and the process is repeated. Previous work has shown that this process results in a substantial enhancement in the number of distinct cDNAs sequences investigated.

Two additional types of mapping information will be critical to the long-term success of primate genomics research. The first is a physical map constructed using radiation hybrid data. Previous genome mapping and sequencing efforts in laboratory or agricultural animals have used radiation hybrid mapping panels to generate physical map data that is independent from, and therefore can be used to confirm and supplement, other types of physical mapping data--including shotgun sequence assembly. While alignment of the nonhuman primate sequence to the human is an excellent strategy for efficient sequence assembly, we anticipate that there will be multiple complex rearrangements that distinguish the rhesus from human. Accurate assembly of the rhesus sequence will benefit from access to radiation hybrid mapping data.

Another type of genomic data critical to the program is genetic linkage data. The baboon genetic linkage map is already described to a resolution of about 330 loci. In order to exploit the potential of nonhuman primates as tools for identifying genes related to disease onset and severity, genetic linkage maps for both rhesus macaques and baboons will be needed. These maps should be improved and extended to include at least 1,000 loci and have average spacing among loci of approximately 2-3 centiMorgans.

Finally, we propose that funding should be made available to allow for genomic sequencing of rhesus macaque. Genomic sequence is critical to identify regulatory sequences and their potential relationship to differential gene expression. Genomic sequence is also necessary to identify fine scale changes in genome organization between rhesus macaques and human that may be relevant to the biology. The rhesus macaque has been chosen as the first nonhuman primate for genomic sequencing for the reasons discussed above. We propose that the method of sequencing should be a whole genome shotgun approach similar to that used by Celera for the human genome. As discussed above, the availability of the full human genome sequence will simplify the assembly of shotgun sequence data from a rhesus macaque. We further suggest that the shotgun sequencing be done to 3-fold redundancy. This level of redundancy will result in approximately 90% complete genomic sequence and represents a reasonable trade-off between expense and completion.

D. Bioinformatics and Data Dissemination

The NHP genome project will generate a substantial amount of valuable mapping and sequence information to be used by the scientific community at large. The ability of the scientific community to make effective use of this information before and after the completion of the genome sequencing effort will depend, in part, on the bioinformatics infrastructure available. This infrastructure includes the databases (existing and new) in which the information is stored and distributed, software tools to access that information, computing hardware to support the system and expert personnel to design and manage the infrastructure. In addition, the data will have maximal impact if resources are put in place to train outside investigators in methods for accessing and analyzing such data.

Several biological databases exist throughout the research community. The relevant biological databases are from two groups: those within the primate community, and those in the general genomics research community. The primate community will be the principal source of specialized information as it relates to primate research, such as background information on species and populations, detailed pedigree information, and phenotype information.

Specialized genomics database systems will also be required to provide annotation for the primate sequences and maps. Identification of structural features within the sequence information will add value to the database. Features such as gene and mRNA organization, protein structure, STS locations, tissue expression profiles, available BAC resources, homology information, motif analysis, and many others are increasingly important to the research community. [The National Center for Biotechnology Information \(NCBI\)](#) and the [European Bioinformatics Institute \(EBI\)](#) have created and/or distribute a number of software tools to aid in this annotation process. However, the level of annotation can differ quite dramatically from center to center. These differences can lead to redundancy in effort and non-uniformity across sequence entries. Coordination between the bioinformatics groups of the funded sequencing centers to standardize the annotation process (or features) would improve the final sequencing product. In addition, interaction between the primate centers is also recommended. As data consumers for this information, they can contribute unique insights to both the annotation and curation process.

To increase the accessibility of NHP genome information to the research community, a centralized portal should be created that contains the NHP Genome Database (NHPGD) and the integrated information discussed above. This reference database will be the primary data for nonhuman primate genomics. Researchers would be able to gather information from one location, rather than from multiple sites under different management schemes. This centralized site will require the appropriate hardware, software and network capabilities in order to serve the community. Existing sequencing centers have the appropriate computing infrastructure to handle this, however it was thought that these centers may not be the best central organizing entities for the NHP genome initiative.

Funding should be secured to support the creation of the NHPGD as well as the informatics necessary to maintain this important scientific resource. This should include additional hardware, software, and qualified personnel for support and development. The individuals involved in the creation and maintenance of the NHPGD may or may not be associated with the existing primate centers. However, they should work closely with both the sequencing centers and the primate centers.

During the workshop, a number of issues were discussed concerning the bioinformatics support for an NHP Genome Initiative. The following sections outline the suggestions from the workshop.

E. Establishment of a Bioinformatics Steering Committee

The ultimate value of the NHP sequence data will be derived from the use of this information to understand biology and disease. This resource should allow investigators to ask critical biological questions and improve

experimental design. The NHPGD will organize a broad scope of information ranging from molecular structures to functions and phenotypes. Therefore, the organization and content of this resource is very important. In order to coordinate this effort, the workshop participants recommended the creation of a steering committee. Members of this committee should come from the funded sequencing centers, the primate centers, the funding agencies, and external members. The external members should come from groups such as the NCBI and EBI. This committee should meet once a year to set milestones, review progress, set priorities, and plan for future needs. The principal function of this committee would be to coordinate the following efforts:

- Planning and creation of a nonhuman primate genome database (NHPGD)
- Coordination of annotation efforts between the sequencing centers
- Coordination of data integration between primate centers
- Sequence annotation
- Planning and supervision of the growth of the database to include other information such as expression array and/or proteomics data

F. Primate Center Involvement

As the major data consumers of this resource, the primate centers are likely to be involved in the planning of bioinformatics initiatives. Layering the existing primate pedigrees, disease models, linkage maps, expression analyses, and other phenotypic data onto the sequence data will be a tremendous asset. In order for the primate centers to be involved, they too will require substantial bioinformatics infrastructure. Funding should be available for this enhancement to the primate centers.

G. Steering Committees for Sequencing and Functional Genomics

We also expect that significant discussion, planning, and oversight will be necessary in relation to the proposed genomic DNA sequencing, cDNA sequencing, genetic and physical mapping, and functional genomics. We recommend that a steering committee be established to provide guidance, planning, and oversight for the sequencing efforts. A separate committee should be established to oversee development of reagents for functional genomics, and implementation of effective functional genomics centers.

Authorship

A committee chosen from among attendees at the January 22-23 meeting wrote this report. This committee included Jeffrey Rogers, Chair (Southwest RPRC); Gary Baskin (Tulane RPRC), Roger Bumgarner (Washington RPRC), Mike Cherry (Stanford University), Scott Hemby (Yerkes RPRC), Lee Hood (Institute of Systems Biology), Jae Jung (New England RPRC), Michael Katze (Washington RPRC), Leslie Lyons (California RPRC), Richard McIndoe (University of Florida), Sergio Ojeda (Oregon RPRC), Gerald Schatten (Oregon RPRC), and David Watkins (Wisconsin RPRC).

A list of those who attended the Workshop is available from Dr. William Morton (Washington RPRC).

Submitted to NCRR February 14, 2001.

Endorsed by the National Advisory Research Resources Council May 17, 2001.