Justification for Whole Genome Sequencing of the Rhesus Macaque (*Macaca mulatta*) Genome

Submitted by Jeffrey Rogers[1], Michael Katze[2],

Roger Bumgarner[2], George Weinstock[3] and Richard Gibbs[3]

[1]Southwest Foundation for Biomedical Research, and Southwest Regional Primate Research Center, San Antonio, TX

[2]Washington Regional Primate Research Center, Seattle, WA

and [3]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX

Correspondence to: Jeffrey Rogers
Dept. of Genetics
Southwest Foundation for Biomedical Research
7620 N.W. Loop 410
San Antonio, TX 78227
jrogers@darwin.sfbr.org

INTRODUCTION

This document summarizes the justification for generating whole genome DNA sequence information for the rhesus macaque (*Macaca mulatta*). We enthusiastically recommend that the National Human Genome Research Institute give the highest level of priority to a large-scale sequencing effort in the rhesus monkey. The scientific justification for and potential impact of the rhesus genomic sequence are outlined in detail below. Overall, the six fundamental characteristics of rhesus monkeys that justify this selection are:

1) Nonhuman primates provide animal models of human disease that are essential to much of biomedical research, and rhesus macaques are the most widely used nonhuman primate. Information from the NIH indicates that in recent years, 60-75% of all nonhuman primates used in NIH funded projects are rhesus monkeys. This species is available in large numbers, and the demand continues to grow. The NIH is supporting the expansion of rhesus breeding programs in order to increase the number of these animals available to researchers. The impact of nonhuman primate genomic data will be greatest if it relates directly to the most utilized model primate.

2) Rhesus monkeys are used in an extraordinarily wide range of biomedical research applications as well as various subfields within basic biology. Due to their close genetic, physiologic, and metabolic similarity to humans, this species serves as an essential research tool in neuroscience, behavioral biology, reproductive physiology, neuroendocrinology, endocrinology, cardiovascular function, pharmacology and many other areas. Consequently, whole genome sequence data will facilitate new and more sophisticated research projects in both existing macaque models of particular human diseases and basic research in more fundamental aspects of mammalian and human biology.

3) By virtue of the way it responds to infection with simian immunodeficiency virus (SIV), this species is widely recognized as the best model of disease progression and pathogenesis in AIDS. Development of new vaccine strategies to fight HIV depend heavily on basic immunology and virology performed using rhesus monkeys, as well as specific experiments using rhesus to test candidate vaccines.

4) As members of the primate superfamily Old World monkeys (Cercopithecoidea), the most recent common ancestor of rhesus and humans lived approximately 25 million years ago. For this reason, and based on a small amount of macaque sequence already available, the overall sequence of rhesus macaques is expected to differ by 5-7.5% from that of human. This level of divergence makes rhesus ideal for a number of molecular comparisons with humans. The substantial similarity between the two species will make the rhesus sequence useful in efforts to identify unknown genes and elucidate various features of the human genome. On the other hand, the overall level of 5-7.5% divergence means that regions of greater than expected conservation between human and rhesus can be used as indicators of previously unrecognized regulatory elements or other sequences in which functional constraints have slowed evolutionary change. Thus, access to whole genome sequence from rhesus will assist in the interpretation of the finished human sequence, including reconstruction of its evolutionary history.

5) The NIH is allocating money to increase the number of rhesus available to investigators because rhesus are widely expected to become more important over time. Given their value for

studies that employ the evolving technologies of gene expression chips or arrays, and proteomics using mass spectrometry, we expect that genomic data for rhesus will create remarkable opportunities for functional genomics. The application of modern proteomics technologies to animal model studies depends on accurate DNA and cDNA sequence for the species under study. The potential impact of high capacity gene expression analysis and state-of-the-art proteomics in studies of an organism so closely related to humans, yet highly amenable to selective breeding, fetal and developmental genetics and experimental manipulation, is extraordinary.

6) Rhesus macaques are closely related in evolutionary and genetic terms to three other widely used laboratory primates (baboons, cynomolgus macaques and pig-tailed macaques). The pairwise inter-species sequence difference among these four species is expected to be 1-2%, which means that whole genome sequence data for rhesus will be tremendously valuable to researchers working on these other species. Access to genomic sequence data from rhesus will create new research opportunities for all four of these commonly studied species.

Humans are members of the Order Primates, and as such, our closest evolutionary relatives are other primate species. This makes primate models of human disease particularly important, as the underlying physiology and metabolism, as well as the genomic structure and content, are more similar to humans than are other mammals. Chimpanzees (*Pan troglodytes*) are the animals most similar to humans in overall DNA sequence, with a difference between the species of approximately 1-1.5% (Stewart and Disotell 1998, Page and Goodman 2001). 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 (*Chlorocebus aethiops*). In contrast, squirrel monkeys, tamarins and marmosets are all New World primates. Like Old World monkeys, New World species are widely used in biomedical research, but they are more distantly related to humans than are the Old World monkeys (e.g. baboons and macaques) or apes. Consequently, there are larger genetic and physiological differences between humans and New World primates than between humans and Old World monkeys and apes. The human evolutionary lineage separated from the ancestors of chimpanzees about 6-7 million years ago (MYA), while the human plus ape lineage diverged from Old World monkeys about 25 MYA (Stewart and Disotell 1998), and from New World monkeys more than 35-40 MYA. In comparison, humans diverged from mice and other non-primate mammals about 65-85 MYA (Kumar and Hedges 1998, Eizirik et al 2001). Thus, humans have been evolving independently of mice for about three times as long as we have been separate from the lineage that includes rhesus monkeys.

A. SPECIFIC BIOLOGICAL RATIONALE

1. *Improving human health.* The choice of rhesus macaques for whole genome sequencing is based on two fundamental ideas: a) our understanding of human biology and disease will benefit tremendously from expanded analysis of the genetic and genomic aspects of nonhuman primate models of disease, and b) the most valuable nonhuman primate for investigating the molecular and genomic aspects of the largest number of human diseases is the rhesus macaque. Primates provide unique models of human disease as a result of their remarkable biological similarities to humans, which include susceptibility and response to infectious disease, neurobiology and

behavioral science, endocrinology, prepubertal and pubertal development, aging, menopause and its consequences, cardiovascular function and many other areas. The selection of rhesus as the optimal nonhuman primate for sequencing is based on the number of individuals available for use, access to large multigeneration pedigrees suitable for pedigree-based genetic analyses and gene mapping, and the breadth of scientific and medical questions that are investigated using this species. We believe that the underlying rationale for sequencing a model organism should be to maximize the amount of information that will be relevant for understanding the molecular and cellular basis of human disease processes. The generation of whole genome sequence for the nonhuman primate most widely used in biomedical research will create the greatest potential for incorporating genomic data and genomic technology into research that will improve human health.

At the present time, rhesus monkeys are the most commonly used nonhuman primate in biomedical research. Current figures from the CRISP database indicate that 60-70% of all NIH-funded grants that involve primates use rhesus macaques. According to the USDA Animal Welfare Report for 2000, about 57,000 primates were used in research that year, and the NIH conservatively estimates that 60-75% of those were rhesus. A literature search in the PubMed database using the combined search terms "rhesus and 2000" generated 935 citations, while a search using "chimpanzee and 2000" produced only 376 and "baboon and 2000" just 387. Figures for other recent years show similar proportions. Clearly, rhesus macaques are central to the biomedical enterprise in this country, and there is substantial opportunity for researchers with a variety of interests and goals to utilize data concerning the rhesus genome within their own research programs. Detailed genomic information would allow investigators to rapidly incorporate state-of-the-art genetics and genomics into nonhuman primate disease models that cannot presently be explored at that level.

In the longer term, the availability of rhesus sequence data will undoubtedly create opportunities for new disease models. The physiological and genetic similarities between humans and rhesus monkeys make these animals outstanding subjects for disease-related research. At present, it is difficult to investigate human-rhesus similarities at the molecular level using high-throughput methods. Access to whole genome sequence information will allow researchers to develop new models of human diseases in rhesus macaques that explore cellular or molecular processes that cannot be examined directly in humans, and are not effectively modeled in non-primate organisms.

It is not possible to list in this document all the ways in which biomedical researchers use and depend upon this species. A few examples must suffice to demonstrate that rhesus are critical to future research progress. While several primate species are used to test candidate vaccines intended to fight AIDS, only macaques are used to study both vaccine strategies and patho-genesis. The rhesus macaque is considered to be the most important animal model of infection and disease progression by the AIDS research community. This includes studies of infection, pathogenesis and treatment of animals infected *in utero* (e.g. Tarantal et al 1999). Studies of viral transmission and replication in fetuses depend entirely on macaque models, as opposed to other primates, some of which can be used to test immunogenicity, but not pathogenesis.

The field of neurobiology is particularly tied to the use of rhesus and other macaques as subjects. A PubMed search using the terms "rhesus and brain" produced 3187 citations, whereas "chimpanzee and brain" yielded just 274. Using the terms "macaque and brain" produced

11,702 citations. Both pharmacology and endocrinology make similar extensive use of rhesus monkeys. This species is often used in analyses of obesity, cardiovascular disease and diabetes (e.g. Hotta et al 2001, Winegar et al 2001). The rhesus macaque is the standard organism of choice in studies of drug addiction, alcoholism and a range of behavioral disorders. Given current national defense concerns, we note that rhesus macaques are used in studies related to protection against anthrax and other aspects of bioterrorism (e.g. Fellows et al 2001). This line of research is certain to expand in coming years. In addition, rhesus are used in programs of gene therapy, as models for the development and testing of vectors, as well as in tests of gene therapy protocols (e.g. Takatoku et al 2001, Lozier et al 2002). The potential applications of rhesus genomic data are remarkably diverse, and touch all aspects of human health.

2. *Informing human biology.* In order for a model organism to make the largest possible contribution to our understanding of human biology, that animal species must be appropriate for the widest possible range of research applications. As discussed above, rhesus monkeys are very well suited to such a role. Studies of the normal, adult physiology of primates can make significant contributions to our understanding of human biology in general and human health specifically. Much of this work is relatively non-invasive and requires little manipulation of individual animals. However, it is also important to note that macaques are appropriate for invasive and terminal studies that are not done on apes. Rhesus are commonly and readily bred and maintained under highly controlled laboratory circumstances, which makes them excellent subjects for well controlled studies of diet, or exposure to chemical or biological insults. For example, rhesus are currently being used in long-term studies of calorie restriction and the physiology of aging (e.g. Lane et al 2000, Edwards et al 2001). Researchers have come to see an increasing number of human diseases as the result of developmental processes, environmental insults or inborn errors that occur or manifest themselves in prenatal or early postnatal development. In order to obtain the maximum benefit from genomic information, the research community will wish to combine genomic data with studies of developmental genetics and age-sensitive developmental processes. The ability to perform invasive and sophisticated experiments on developing fetuses or growing juveniles will maximize the impact of large-scale DNA sequence data. The application of genomic information and technology to such studies in animals so similar to humans will revolutionize our ability to investigate fundamental questions of human biology, and encourage innovative use of large pedigrees of rhesus macaques.

Some traits will be more similar between humans and apes such as chimpanzees than between humans and rhesus. However, the value of a genome sequence is much higher for an organism that can subsequently be used in a wide range of experiments. At present, experimental utilization of chimpanzees and other apes is limited by cost, availability and ethical concerns. This situation is unlikely to change soon, as there is a breeding moratorium in place for all NIH owned chimpanzees, and the federal government is currently considering placing some federally supported apes into a sanctuary system, outside normal biomedical institutions, to reduce cost (Dove 2000; Laboratory Primate Newsletter 40:21, 2001).

Given the large pedigrees of rhesus monkeys that are now available, there is significant opportunity to study gene by environment interaction in the control of phenotypes. This requires large numbers of genealogically related individuals, all examined using the same experimental protocol. This approach has been successful using pedigreed baboons (Mahaney et al 1999). Over 700 genetically characterized baboons were each fed two controlled diets, and blood cholesterol levels were studied as a function of both diet and genetics. Linkage mapping located

a chromosomal segment containing a gene (or genes) that influence response to dietary cholesterol. Overall, nine LOD scores over 3.0 have been obtained in this pedigree for traits related to lipids and cholesterol, hypertension and obesity (e.g. Kammerer et al 2001, Martin et al 2001). This degree of experimental control is simply not possible in human families, and illustrates the unique opportunities that Old World monkeys provide for investigating gene by environment interaction. The expanding pedigrees of NIH supported rhesus macaques will make this type of analysis increasingly feasible in this species, and whole genome sequence would dramatically improve the power of such approaches.

3. *Informing the human sequence.* As discussed above, paleontological information, plus a small amount of genomic sequence available from rhesus, suggests that the overall DNA sequence of rhesus is about 5-7.5% different from the human sequence (Stewart and Disotell 1998, Page and Goodman 2001). This level of divergence is highly advantageous for comparative genomics; the rhesus sequence will provide many opportunities for studies that enhance our understanding of the human sequence. For example, local regions of DNA that regulate transcription are expected to be conserved across many mammals. When two divergent sequences are compared, non-coding regions that have higher than average similarity are likely candidates for regulatory sequences. The anticipated divergence of 5-7.5% between rhesus and humans provides opportunity to look for small regions of greater than average nucleotide identity. If the sequence that is compared to human is too similar (e.g. 1-2% divergent), then conservation due to significant biological function and selective constraint will be masked. While showing substantial divergence from humans, rhesus monkeys are close enough in evolutionary and genetic terms to share the vast majority of functional genes. Genes that are difficult to recognize in the human sequence using current algorithms may be more easily identifiable in the rhesus. Searching for functional genes by combining human sequence with homologous sequence that is about 6% different, but which shares essentially all coding elements, is likely to be more informative than searching either sequence alone. Researchers are already using the small amount of monkey sequence currently public in this manner (Osada et al 2002).

4. *Providing connection between sequences of non-human organisms and humans.* The close evolutionary relationship of rhesus macaques to humans makes it possible to use the two sequences together to represent primate genomes in evolutionary analyses. This will help indicate which aspects of the human genome are shared among various primates, and which aspects are of more recent origin (i.e. have evolved in the last 25 million years). The second perspective on the primate genome provided by a rhesus sequence will make comparisons to mouse, rat and other mammals more meaningful. Mice and humans diverged about 65-85 million years ago (Eizirik et al 2001, Kumar and Hedges 1998). Once both the human and mouse sequences are completed, there will undoubtedly be many detailed comparisons done to reconstruct and interpret the history and content of the mammalian genome. Differences found between mouse and human must be ascribed to mutations that occurred either in the rodent or human lineages. Rhesus macaques have shared with humans about two-thirds of the evolutionary history since primates diverged from rodents. Access to rhesus sequence will facilitate interpretation of the history of the human genome by helping to date mutational events and place them onto evolutionary lineages.

5. *Expand understanding of basic biological processes relevant to human health.* As discussed above, rhesus macaques are used extensively in a wide range of research relevant to human health. However, sequence information can greatly accelerate ongoing research in a

number of ways. In particular, sequence data will enable investigators to pursue functional genomics in this species. For example, the most powerful proteomics analysis methodologies make use of mass spectroscopy combined with sequence data to identify proteins. Rhesus sequence is sufficiently distinct from human sequence, that modern day proteomics methods are not applicable without sequence data from the macaque species itself. Similarly, expression array analysis will be greatly enhanced if rhesus sequence data is available. While human arrays can be and have been used for studies with nonhuman primate samples, we anticipate that there will be some sequences that are present in rhesus which will not be represented on human arrays. In addition, Feldman, Katze and Bumgarner have recently sequenced a number of EST's from rhesus monkeys (see GenBank accession numbers BM423011- BM423313). The mean value of sequence similarity between human and rhesus is about 95%, but there is a long tail to the distribution, consisting of sequences with much lower similarity. Unpublished data from the Katze lab suggests these low similarity sequences produce significantly lower signal when hybridized to arrays made with human cDNAs, thus reducing the value of the quantitative expression array data. Furthermore, 3' UTR segments are often used in expression array assays, because they differ among members of a gene family. 3' UTR sequences are even more different between humans and rhesus than are coding segments. Expression arrays made from rhesus sequences would avoid this problem and produce better results.

We feel that functional genomics applied to rhesus and other closely related primates (e.g. baboons and other macaques) can make outstanding contributions to many subfields within biomedicine and human biology. Clearly, it is unethical and/or impractical to directly investigate gene expression or protein composition within cells of the developing human embryo or fetus. With appropriate IACUC approval, fetal or embryonic tissue from any developmental stage and any organ system can be obtained from rhesus. With the growing interest in stem cell technology and the desire to manipulate stem cells, knowledge of the genetics and genomics of primate fetal development will prove invaluable to human stem cell research. Similar arguments can be made concerning pharmacogenomics. Either young or adult monkeys can be challenged with experimental or approved pharmaceuticals to generate detailed profiles of cellular responses to drugs.

6. *Provide new surrogate systems for human experimentation such as new disease models.*
As described above under sections 1, 2 and 5, rhesus macaques already provide a wide range of experimental models and surrogate systems for understanding human biology. We expect rhesus genomic sequence to encourage and facilitate the development of new disease models, and new experimental strategies for existing models.

7. *Facilitate ability to perform direct genetics or positional mapping.* A genetic linkage map has already been developed for a close relative of rhesus (the baboon, Rogers et al 2000) and a similar map is under development for rhesus (NCRR R01 RR15383, J. Rogers, P.I.). A 10 centiMorgan map of the entire rhesus genome will be available in about 24 months, and this will create new opportunities to use large multi-generation pedigrees of rhesus for gene mapping and gene identification studies. Analyses in baboons have already demonstrated the feasibility and value of quantitative trait linkage mapping in large pedigrees of nonhuman primates (e.g. Martin et al 2001, Kammerer et al 2001). Positional cloning studies are possible without the whole genome DNA sequence for rhesus, but the pace of progress when moving from initial QTL linkage result to identification of functional gene and mutation will be dramatically accelerated in macaques as well as in closely related baboons if the rhesus genome is fully available. Furthermore, several baboon QTLs have been found in regions where intra-chromosomal

inversions or rearrangements make locus order different in humans and baboons. Linkage analysis alone is not sufficiently precise to clarify the exact nature or boundaries of these rearrangements. As a result, physical maps, or better yet sequence, data for rhesus (or baboon) are critical for the rapid identification of the genes and mutations that produce the positive LOD scores obtained in positional mapping studies.

8. *Expand our understanding of evolution in general and human evolution in particular.* As an evolutionary lineage, primates separated from other mammals about 65-85 million years ago (Eizirik et al 2001, Kumar and Hedges 1998). The ancestors of rhesus macaques (and other Old World monkeys) separated from human (and other hominoid) ancestors about 25 million years ago (Stewart and Disotell 1998). Thus, much of the evolutionary history of primates is shared by the genomes of rhesus and humans. On the other hand, the rhesus and human lineages have followed different trajectories for about one-third of primate history. Comparison of whole genomic content and sequence between these two representatives of the Order Primates will generate many opportunities for the analysis of individual genes, gene families, families of repetitive elements and other components of the genome.

B. STRATEGIC ISSUES

1. *Demand for new sequence data.* There is tremendous demand for DNA sequence information concerning rhesus macaques. We include a letter from the Directors of all eight NIH-funded regional primate research centers supporting this effort. These centers serve as resources for investigators in a large number of fields across the nation. We have received 27 additional letters from other researchers expressing support for this recommendation. These letters are available to NHGRI upon request. In addition, rhesus macaques were the consensus choice for intensive genomic studies among participants in the Primate Genomics Workshop held in January 2001 in Seattle. This workshop was jointly sponsored by the University of Washington Regional Primate Research Center and the National Center for Research Resources. Investigators from a wide variety of institutions met to discuss the current state of primate genomics and how progress in this field could be encouraged. An Executive Committee was chosen and given the task of writing a summary report for Dr. Judith Vaitukaitis, Director of NCRR. This Summary Report was submitted to Dr. Vaitukaitis in February 2001, and was later endorsed by the National Advisory Council of NCRR. The text of the report is available at http://www.ncrr.nih.gov/compmed/primate_genomics20010606. The major recommendations are that intensive genomic analyses should be pursued in a number of primate species, and that the largest effort (including whole genome sequencing) should be directed at rhesus macaques.

As discussed above, most of the primates used in biomedical research in the U.S. are rhesus monkeys. In addition, three other commonly used species (baboons, *Papio hamadryas*; cynomolgus macaques, *Macaca fascicularis* and pig-tailed macaques, *Macaca nemestrina*) are all closely related to rhesus macaque. All species of macaques will be no more than about 1% different in genomic sequence, and baboons will be 1-1.5% different from rhesus (Stewart and Disotell 1998, Page and Goodman 2001). Access to extensive genomic sequence from rhesus will significantly benefit research using all these species. Sequence from chimpanzee would not provide any additional information beyond that derived from access to the available human sequence because the evolutionary distance, and sequence similarity, between humans and rhesus (or baboons) is the same as the distance between chimpanzees and rhesus (or baboon).

At present, there is very little data in GenBank for nonhuman primates and almost no data for rhesus macaque. There are now approximately 2000 GenBank entries for rhesus macaque. Since many of these entries are redundant, we estimate that approximately 450-500 unique macaque genes are represented in GenBank. Despite this, utilization of rhesus by NIH-funded investigators is very high. A search of the CRISP database indicates that there are 264 R01 grants currently active that use rhesus macaques, and the need for rhesus monkeys is growing. The National Center for Research Resources has recently funded six new rhesus colonies, with the expressed goal of increasing the availability of rhesus macaques to NIH-funded researchers.

2. *Suitability of organism for experimentation.* As the material above demonstrates, rhesus macaques are highly amendable to experimentation and investigation. This species is available in large numbers from a variety of institutions and breeding operations, and is already used in 60-75% of NIH funded projects that utilize nonhuman primates. In addition to the research described above, rhesus are also used in more novel and ground-breaking research. The first nonhuman primate expressing an exogenous gene has been produced (Chan et al 2001), and this animal was a rhesus monkey. While routine production of transgenic primates is not yet possible, we expect transgenic procedures to be developed and applied to both rhesus monkeys and baboons. Stem cell research will also exploit the strengths of rhesus macaque models over the coming years. A search of the PubMed database for "stem cell and rhesus" generated 210 citations. In comparison,using "stem cell and chimpanzee," the PubMed database found 31 papers, only 5 of which are more recent than 1994.

3. *Rationale for complete sequence.* The biomedical research community will wish to exploit nonhuman primate genomic data for studies of gene regulation and the mechanisms of transcriptional control. To fully exploit the strengths of rhesus and closely related species for such studies, it will be necessary to have as much of the full genomic sequence as possible. As discussed above, new expression array methods that allow investigators to monitor thousands of genes simultaneously will provide opportunities to explore gene regulation in various tissues and at different stages of development, thus creating opportunities to examine the molecular basis of gene regulation at a level of detail not currently possible in any primate, including humans.

4. *Cost of sequencing and readiness of DNA. Quality of sequencing product and strategy.* The genome size for rhesus macaques is expected to be very similar to that of humans. A high-quality BAC library has been produced by Dr. Pieter deJong. The library is being used by several researchers (see below). Two different restriction enzymes (EcoRI and MboI) were used to make the library, and the average insert size is 160kb. As described in detail below, we urge the NHGRI to approve sequencing of this library to an overall coverage of 5-fold redundancy.

5. *Other partial support*. At the present time, no additional support for a rhesus sequencing project has been obtained from NIH institutes or other sources. However, a large number of the categorical institutes within NIH support projects using rhesus. We anticipate that several components of NIH may be willing to invest in this important program.

C. COLLABORATION WITH BAYLOR GENOME SEQUENCING CENTER

Dr. Richard Gibbs and Dr. George Weinstock of the Baylor Genome Sequencing Center (Houston, TX) have expressed strong interest in performing the whole genome DNA sequencing of the rhesus macaque genome at their center (see attached letter). The other authors of this

"white-paper" have visited the Baylor sequencing center and discussed details of the project. We propose sequencing the entire genome to a five-fold average coverage. In addition, we propose that support be allocated for more complete coverage and finishing of sequence for about 500 megabases within the rhesus genome. The specific regions (500MB in total) to receive this additional effort would be chosen at a later time based on perceived scientific interest and importance. For example, given the significance of rhesus macaques for AIDS research and immunology, and for neurobiology, it might be valuable to produce "finished" sequence for the MHC region of the rhesus genome, and for one or more regions that contain a high density of medically significant genes expressed in the nervous system.

The strategy for sequencing of the rhesus macaque genome is guided by the observation that, although conservation between human and macaque in coding regions averages 92-95%, there are considerable regions with less conservation. Moreover non-coding regions may be as low as 85% conserved. Thus any strategy of low coverage sequencing followed by comparison with the human sequence will certainly miss many features of interest, i.e. the more divergent regions. The strategy will thus aim for a higher coverage (5-6x) of the genome. This will employ a mixed approach of low coverage (1-2x coverage BAC skims) sequencing of a minimal tiling path (MTP) of BAC clones with the rest of the sequence coming from whole genome shotgun reads.

There are at present two BAC libraries that have been prepared by Pieter de Jong (from EcoRI- and MboI-generated fragment sets, average insert size of 160kb). At the BCM-HGSC ten of these clones are currently being sequenced to determine whether any unforeseen issues exist. There are no BAC end sequences (BES) published, nor is a fingerprint map available. However, Dr. Shaying Zhao at The Institute for Genomic Research (TIGR) has sequenced both ends of approximately 4600 BAC clones from the deJong library, and finds that most of these sequences can be readily aligned to the human sequence (Zhao, pers. comm., letter available on request). This indicates that the BAC library can provide high-quality DNA. BAC end sequences from ~5000 clones would not be sufficient to generate a MTP without additional effort. A MTP can be derived at the BCM-HGSC by generating BES from an additional 70-75,000 clones, corresponding to 4x clone coverage of the genome, and mapping these to the human genome. The high conservation between human and macaque will allow us to use the human sequence as a scaffold to order and orient the macaque BES. A MTP will be picked from this map, corresponding to about 25,000 clones (about 30% overlap between clones). The clone overlaps will initially be verified by restriction fragment fingerprinting. Clone gaps will be filled by screening the libraries with oligo probes based on BES or BAC skims (to be done later).

The 25,000 BACs will be sequenced using a pooled clone array (CAPSS) approach to minimize the number of BAC DNA preparations and shotgun libraries required (Cai et al 2001). In the extreme form of this method, the BACs will be arranged in a 160x160 array and the 320 rows and columns will be pooled separately. DNA preparations and shotgun libraries will be prepared from these pools and sequenced to an average coverage of 1-2x per BAC. The sequences from each row and column will be mixed in each pairwise combination. The mixtures of reads will be compared to the human genome and row and column reads that map near each other (within a BAC insert of each other) will be judged to be from the BAC at the intersection of the row and column. The mixed row-column reads will also be co-assembled and contigs with both row and column reads will be judged to come from the BAC at the intersection of the row and column. Using these two criteria, reads will be assigned to each BAC. This method reduces the number of DNA preparations and shotgun libraries from 25,000 to 320. A more conservative

approach, that may be more manageable technically, would use 60 arrays of 20x20 clones, requiring 2400 DNA preparations and libraries, which is still a 10-fold reduction. Experiments underway will determine what is a reasonable sized array in terms of laboratory manipulations.

Whole genome shotgun reads will come from 3kb, 10kb, and 40kb libraries. All sequencing will be in plasmids and end pairing will be maintained at high (>90%) fidelity. The whole genome shotgun reads will be binned into appropriate BACs and the genome will be assembled using the ATLAS whole genome assembly software developed at BCM-HGSC for the rat genome project. The final assembly will also take advantage of the human genome sequence to provide additional validation. In general, all of the methodology employed in this project is being used in the rat genome project and will be proven technology.

In addition to the construction of the draft DNA sequence, two other activities are of high interest. The first is finishing of regions of interest. As much as 500MB of the macaque genome may be of high interest for human disease studies and sufficiently different from the human genome to require finishing for interpretation. The second additional activity is sequencing of full-length cDNAs from tissues that are not readily available from humans (e.g. embryonic tissues) and represent one of the key advantages of a nonhuman primate system. The BCM-HGSC currently has a high-throughput full-length cDNA sequencing pipeline that would be appropriate for this.

The entire project would require approximately 35 million successful reads which could be performed at the BCM-HGSC in about two years. However, it is likely that this time would be reduced by sharing responsibility for the WGS component with another group. The BAC DNA preparations and libraries would take one year if all 25,000 BACs were to be analyzed individually, and considerably less time using the pooling strategy. Thus the entire project could be completed in two years.

D. FINAL COMMENTS

We enthusiastically recommend that the rhesus macaque be chosen for high priority whole genome sequencing. This species is the most important nonhuman primate in biomedical research. The impact of rhesus genome sequence on studies of human health and disease will be diverse and highly significant. As we have described above, the fields of neurobiology, AIDS research, reproductive biology, endocrinology, cardiovascular disease, diabetes, obesity and many others will benefit tremendously from this additional information. Functional genomics and the interpretation of the human genomic sequence will also see substantial benefits. We strongly believe that the rhesus macaque is the correct choice for the first nonhuman primate to be sequenced. While the general field of human biology, and human evolutionary genetics in particular, would benefit from full genomic sequence of other primate species such as chimpanzees and baboons, the material above demonstrates that the species with the greatest potential impact on our overall understanding of human disease, and therefore the one with the highest priority for immediate sequencing, is the rhesus macaque.

Literature Cited

Cai, W.-W., R. Chen, R.A. Gibbs and A. Bradley (2001) "A clone-array pooled shotgun strategy for sequencing large genomes" **Genome Res.** 11:1619-1623.

Chan, A.W., K.Y. Chong, C. Martinovich, C. Simerly and G. Schatten. (2001) "Transgenic monkeys produced by retroviral gene transfer into mature oocytes" **Science** 291:309-312.

Dove, A. (2000) "New bill mandates sanctuary system for retired chimps" **Nat. Medicine** 6:9.

Edwards, I.J., L.L. Rudel, J.G. Terry, J.W. Kemnitz, R. Weindruch, D.J. Zaccaro and W.T. Cefalu (2001) "Caloric restriction lowers plasma lipoprotein (a) in male but not female rhesus monkeys" **Exp. Gerontol.** 36: 1413-1418.

Eizirik, E., W.J. Murphy and S.J. O'Brien (2001) "Molecular dating and biogeography of the early placental mammal radiation" **J. Heredity** 92:212-219.

Fellows, P.F., M.K. Linscott, B.E. Ivins, M. Pitt, C. Rossi, P.H. Gibbs and A.M. Friedlander (2001) "Efficacy of a human anthrax vaccine in guinea pigs, rabbits and rhesus macaques against challenge by Bacillus anthracis isolates of diverse geographic origin" **Vaccine** 19:3241-3247.

Hotta K, Funahashi T, Bodkin NL, Ortmeyer HK, Arita Y, Hansen BC, and Matsuzawa Y. 2001. Circulating concentrations of the adipocyte protein adiponectin are decreased in parallel with reduced insulin sensitivity during the progression to type 2 diabetes in rhesus monkeys. **Diabetes** 50:1126-1133.

Kammerer, C.M., L.A. Cox, M.C. Mahaney, J. Rogers and R.E. Shade (2001) "Sodium-lithium countertransport activity is linked to chromosome 5 in baboons" **Hypertension** 37:398-402.

Kumar, S. and S.B. Hedges (1998) "A timescale for vertebrate evolution" **Nature** 392: 917-920.

Lane, M.A., E.M. Tilmont, H. DeAngelis, A. Handy, D.K. Ingram, J.W. Kemnitz and G.S. Roth (2000) "Short-term calorie restriction improves disease-related markers in older male rhesus monkeys (Macaca mulatta)" **Mech. Ageing Dev.** 112: 185-196.

Lozier, J.N., G. Casko, T.H. Mondoro, D.M. Krizek, M.E. Metzger, R. Costello, J.G. Vostal, M.E. Rick, R.E. Donahue and R.A. Morgan (2002) "Toxicity of a first-generation adenoviral vector in rhesus macaques" **Hum. Gene Ther.** 13: 113-124.

Mahaney, M.C., J. Blangero, D.L. Rainwater, G.E. Mott, A.G. Comuzzie, J.W. MacCluer and J.L. VandeBerg (1999) "Pleiotropy and genotype by diet interaction in a baboon model for atherosclerosis. **Arterioscler. Thromb. Vasc. Biol.** 19: 1134-1141.

Martin, L.J., J. Blangero, J. Rogers, M.C. Mahaney, J.E. Hixson, K.D. Carey, P.A. Morin and A.G. Comuzzie (2001) "A quantitative trait locus influencing estrogen maps to a region homologous to human chromosome 20" **Physiol. Genomics** 5:75-80.

Osada, N., M. Hida, J. Kusuda, R. Tanuma, M. Hirada, K. Terao, Y. Suzuki, S. Sugano and K. Hashimoto (2002) "Prediction of unidentified human genes on the basis of sequence similarity to novel cDNAs from cynomolgus monkey brain" **Genome Biol.** 3:RESEARCH0006.

Page, S.L. and M. Goodman (2001) "Catarrhine phylogeny: Noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade" **Mol. Phylogenet. Evol.** 18:14-25.

Rogers, J., M.C. Mahaney, S.M. Witte, S. Nair, D. Newman, S. Wedel, L.A. Rodriguez, K.S. Rice, S.H. Slifer, A. Perelygin, M. Slifer, P. Palladino-Negro, T. Newman, K. Chambers, G. Joslyn, P. Parry and P.A. Morin (2000) "A genetic linkage map of the baboon *(Papio hamadryas)* genome based on human microsatellite polymorphisms" **Genomics** 67: 237-247.

Stewart, C.-B. and T.R. Disotell (1998) "Primate evolution – in and out of Africa" **Current Biology** 8:R582-R588.

Takatoku, M., S. Sellers, B.A. Agricola, M.E. Metzer, I. Kato, R.E. Donahue and C.E. Dunbar (2001) "Avoidance of stimulation improves engraftment of cultured and retrovirally transduced hematopoietic cells in primates" **J. Clin. Invest.** 108:447-455.

Tarantal, A.F., M.L. Marthas, J.P. Shaw, K. Cundy and N. Bischofberger (1999) "Administration of 9-2-R-(phosphonomethoxy)propyladenine (PMPA) to gravid and infant rhesus macaques (Macaca mulatta): safety and efficacy studies" **J. Acquir. Immun. Defic. Syndr. Hum. Retrovirol.** 20:323-333.

Winegar DA, Brown PJ, Wilkison WO, Lewis MC, Ott RJ, Tong WQ, Brown HR, Lehmann JM, Kliewer SA, Plunket KD, Way JM, Bodkin NL, and Hansen BC. 2001. Effects of fenofibrate on lipid parameters in obese rhesus monkeys. **J Lipid Res** 42:1543-1551.