Report from the Annotating the Human Genome Working Group
April 2007

This report is divided into five sections:

1. Overview
2. Part 1: Biomedical Justifications
3. Part 2: Sequencing Primate Genomes to Annotate the Human Genome
4. Part 3: Other Justifications
5. Appendices

Overview

The Working Group identified multiple broad categories of rationales for the additional sequencing of primate genomes, divided into the three inter-related below. Note that a subset of these, indicated with *'s, reflect rationales that were subjected to a greater amount of discussion and were regarded, in general, as the higher-priority reasons for additional primate-genome sequencing at this time.

1. Biomedical Justifications

A*. Enhancing the value of individual biomedical models and establishing new models for specific human diseases
B*. Understanding the basis of disease-phenotype differences between species (within and among lineages)
C. Enhancing the ability to use primates as models for understanding embryonic development

2. Sequencing Primate Genomes to Annotate the Human Genome

D*. Finding primate-specific genomic elements
E*. Finding genomic regions that have been subject to recent evolutionary selection within primates (or sublineages therein)
F. Understanding rapidly-evolving genomic sequences within the primate lineage (at multiple scales, and including copy-number variants, expansions of gene families, and other sequence elements)
G. Reconstructing evolutionary changes of the genome in the hominid lineage
H. Understanding the basis of non-disease-related phenotypic differences between species (within and among lineages)

3. Other

I*. Understanding within-species variation (may be extendable to variation between closely-related species); related to basic population genomics and basis of disease phenotype
J. Using primate sequence to fill in gaps in the human reference sequence

Each of these three categories is detailed in a separate section below, in each case with individual primate species proposed. A summary of the prioritized list of primate species being proposed here is provided in **Appendix 2**. This list includes 12 species, requiring roughly 160 Gb of whole-genome shotgun sequencing capacity plus some targeted sequence refinement.

Note 1: Sequence Quality
In earlier proposals for primate-genome sequencing, this Working Group proposed (successfully) the generation of high-quality primate-genome sequence (e.g., 6-7-fold shotgun coverage with the directed finishing of ~1000 BACs to resolve regions of weak quality and/or significant duplicated regions). We propose the same for all of the primate species nominated herein (**unless otherwise noted**). Without fully reiterating our previous arguments, there are several key reasons for generating high-quality sequences for primate genomes:

- Differences rather than similarities represent the important findings. It is critical to detect subtle differences in genes or their regulatory regions that could lead to phenotypic differences, and (importantly) to know that such differences are due to biology and not to poor sequence quality.
- Large-scale genome rearrangements need to be resolved, including duplications and other structural rearrangements.

Even when the major rationale for sequencing a primate species is based on biomedical significance and its role as a research model, a high-quality genome sequence is needed to serve as a reliable resource for the research community.

Note 2: cDNA resources
cDNA resources have proven an invaluable addition to many genome sequencing projects, mainly for aiding in annotation of genes, but also as a resource where there are significant user communities. In addition, new technologies make them relatively inexpensive to sequence. (However, we note that, for some of the species proposed here, obtaining tissues will be difficult). The working group thus proposes that cDNA resources (specifically, 100,000-200,000 ESTs and possibly some full length clones) be established for each of the species nominated here, where they do not already exist.

Note 3: New DNA Sequencing Technologies
The Working Group discussed the potential suitability of new DNA sequencing technologies for generating primate-genome sequences. Because the current proposal will require a considerable amount of current sequencing capacity, cost is a significant consideration. In general (and relevant to what is proposed here), the new short-read sequencing technologies are currently best-developed for cDNA/EST projects, finding SNPs, sequencing small genomes (e.g., bacterial), and potentially for surveying structural variation; where such studies are proposed, it would make sense to use (or to at least pilot) these new technologies. But most of the sequencing proposed here will involve de novo assembly of large genomes, where it remains unestablished how to best utilize the

new short-read technologies. However, this is a rapidly changing area, and one that the sequencing centers and NHGRI are closely monitoring. As suitable applications with these new technologies emerge for primate-genome sequencing, they should certainly be piloted and carefully evaluated.

Part 1: Biomedical Justifications

There are a number of different primate species that are important as animal models in the study of human health and disease. Consequently, such primate species would be appropriate candidates for whole-genome sequencing based on a focused biomedical rationale. Among these, there is a smaller set of primates that are widely used and hence critically important to research progress. We feel that priority should be given to those species for which whole-genome sequencing will produce the greatest overall benefit to research progress related to the causes and treatment of human disease.

Below we divided our list of high-priority species into three groups: Immediate Priority, High Priority, and Strong Justification. Species are further prioritized within each group.

Immediate Priority (should begin as soon as possible)

1) Baboons (*Papio hamadryas*)
2) Cynomolgus macaque (*Macaca fascicularis*)
3) Squirrel monkey (*Saimiri* sp.)

High Priority (will have significant biomedical impact in many research areas)

4) Owl monkey (*Aotus* sp.)
5) Vervet or African green monkey (*Chlorocebus aethiops*)
6) Chinese rhesus macaque (*Macaca mulatta lasiota and M.m. sanctijohannis*)

Strong Justification (will provide unique, important information in several areas)

7) Pigtail macaque (*Macaca nemestrina*)
8) Sooty mangabey (*Cercocebus atys*)

Immediate Priority

1) Baboons (*Papio hamadryas*)

Baboons (*Papio hamadryas*) are among the most widely used primate species in biomedical research. They are highly adaptable, large-bodied omnivorous Old World monkeys that reside widely in sub-Saharan Africa and the Arabian Peninsula. Baboons are members of the subfamily Cercopithecinae, more closely related to macaques than to African green monkeys, but slightly more distant phylogenetically from macaques than humans are from chimpanzees. The species has five readily distinguished subspecies that differ in body size, coat color, morphology (e.g. skull shape), physiology (e.g. lipoprotein profiles), and social behavior.

A recent survey of the eight NIH primate research centers indicates that baboons rank third in terms of the numbers of animals used, behind only the rhesus macaque and the cynomolgus macaque. A literature search in PubMed for the year 2006 (using the search term “baboon”) produced 235 publications. Fifty papers randomly selected from that list of 235 had first authors from 37 different institutions, documenting broad use of these animals across a significant number of laboratories and institutions. Baboons are available in the U.S. in large numbers, with costs per animal much less than many other primate species. Significant breeding colonies exist at several institutions, including the University of Oklahoma Health Sciences Center, the Yerkes National Primate Research Center (NPRC), and the Washington NPRC. The largest U.S. breeding colony of baboons is at the Southwest NPRC.

Baboons are used as animal models for a broad range of diseases and disease-related biological processes. This includes models of infectious disease, metabolic diseases, neurobiological processes, and others. One unique use is in a major program studying bronchopulmonary dysplasia and other clinical aspects of premature birth. This model depends on the production of viable premature infants at well-defined stages of development. The unique neonates produced are used by several different NIH-funded projects examining different aspects of pre-term birth and its consequences. Due to their large body size (adults are more than twice as large as adult rhesus macaques), baboons are favored by investigators studying organ transplantation, xenotransplantation, the development and testing of gene therapy vectors, and other topics requiring surgical manipulation of animals. For the same reason, baboons are widely used in studies of structural and functional brain imaging. Other important uses involve studies of infectious diseases, including schistosomiasis, Chagas disease, and herpes that infect tens of millions of people across the world. Baboons are also frequently used in analyses related to epilepsy, endometriosis, and various aspects of reproductive biology.

One of the most significant uses of this species has been in the genetic analysis of risk factors for common metabolic diseases, especially diabetes, atherosclerosis, osteoporosis, obesity, and hypertension. Some baboons spontaneously develop diabetes, obesity, hypertension, or osteopenia, but this is variable and related to individual genetic differences. These metabolic diseases account for substantial morbidity and mortality in the U.S. and around the world. Baboons are used for investigation of the underlying pathophysiology as well as genetic risk.

We anticipate that access to the complete genome sequence of baboons would lead directly to an improved understanding of the biological problems listed above. Whole-genome sequence would facilitate more precise and comprehensive investigation of gene expression in the baboon, and would eventually permit intensive proteomic analysis. One of the immediate impacts of the baboon genome sequence would be acceleration of progress in the identification of genes that affect disease-related phenotypes. Using the available multigenerational pedigrees and whole-genome linkage map, researchers have mapped more than 20 quantitative trait loci (QTLs) that influence risk factors for metabolic diseases such as cholesterol levels, bone density, and adipocyte-related phenotypes. Investigators have also begun to generate information about expression QTLs. The complete baboon sequence would accelerate identification of the underlying functional mutations, as well as identify gene copy-number variation and other functionally significant genomic features that are currently undescribed.

Given the significant background information available for baboons, it is likely that a whole-genome sequence would create opportunities to expand existing disease models (such as those for fetal development and preterm birth, immunology and infectious disease, and neurobiology) and to develop novel models based on new information concerning baboon-human similarities. The substantial genetic, morphological, and behavioral variation found within this single species also means that genomic analysis will contribute to our understanding of genotype-phenotype relationships in a wide range of anatomical, physiological, or developmental systems.

Finally, there is further justification for sequencing the genome of a second Old World monkey. We expect at least as many differences in gene content, segmental duplications, and other significant features to exist between baboons and rhesus monkeys as are observed between humans and chimpanzees. This is based on the phylogeny and available estimates of single-copy DNA sequence divergence (1.7% substitutions, 0.24% indels between baboon and rhesus). Knowledge of the baboon DNA sequence will improve our understanding of primate-genome evolution and the reconstruction of the ancestral Old World monkey-hominoid genome by adding a second Old World monkey dataset to all human-hominoid-cercopithecoid analyses. (However, it will not help with this goal as much as a more distantly-related old-world monkey).

2) Cynomolgus Macaques (*Macaca fascicularis*)

Cynomolgus macaques (*Macaca fascicularis*) are a critically important nonhuman primate in biomedical research, used in large numbers by investigators in pharmacology, breast, uterine and other cancers, diabetes, cardiovascular disease, and other fields. Cynomolgus macaques are closely related to rhesus macaques (*Macaca mulatta*). While clearly distinct in body size (cynomolgus macaques are smaller), physiology, and susceptibility to infectious diseases, these two species can form reproductively viable hybrids in the wild. Cynomolgus macaques are widely distributed in Malaysia, Indonesia and the Philippines. They inhabit tropical forest, but are quite tolerant of captive housing and diets. There are several major breeding colonies in the U.S., but large numbers of these animals can be imported from Indonesia and other sources, making them readily available to laboratories at much lower cost than rhesus monkeys. *Macaca fascicularis* ranks second only to the rhesus macaque in numbers of individuals available in the NIH primate research center program. A PubMed search using the search term "cynomolgus" yielded 264 papers published in the year 2006 alone.

Because they are relatively small in comparison to other Old World monkey laboratory species, cynomolgus macaques have been widely used in drug development, drug testing, and toxicology. They also show a valuable pattern of lipoprotein and other cardiovascular responses to dietary cholesterol, dietary fat, and other risk factors. This species is highly responsive to stress, and thus has frequently been used to test interactions between stress and other cardiovascular risk factors. Diabetes is also characteristic of a significant fraction of cynomolgus monkeys, and this has been exploited by numerous investigators. Estrogen physiology, the effects of psychosocial stress on reproductive function, and the interactions of estrogens with dietary and other risk factors for disease have all been studied in this species. Finally, cynomolgus macaques have received significant attention as models for drug abuse, alcohol abuse, depression, and other psychiatric/psychological disorders.

The impact of whole-genome sequencing of the cynomolgus macaque would be very significant. Accurate information about gene content and the specific sequences for cynomolgus genes would improve opportunities for studies of pharmacology and pharmacogenomics, responses to infectious disease, and metabolic disorders such as diabetes and osteoporosis. While cynomolgus macaques are closely related to rhesus, differences in gene number and sequences mean that quantitative studies of gene expression may benefit from species-specific genomic data. Given their small size, lower cost, and more convenient handling, use of this species would likely increase, if sophisticated analyses and tools based on state-of-the-art genomic data were available. Expansion of the diversity of primate species used for biomedical research is a goal within the NIH, given the high cost of overdependence on rhesus macaques.

Genetic, morphological, and physiological differences among populations of cynomolgus macaques are significant, and becoming well-known. This species is well-suited to detailed studies of genomic variation, the phenotypic consequences of genetic variation, and the consequences of selection on the primate genome. While closely related to the rhesus macaque, there are biologically meaningful differences in physiology, especially susceptibility to infectious disease. Therefore, detailed comparison of these two species of macaques at the physiological, cellular, and genomic levels would provide insight into both mechanisms of disease and primate responses to selection and the neutral forces of evolution.

3) Squirrel monkey (*Saimiri* sp.)

The squirrel monkey (*Saimiri* sp.) is a moderately sized, neotropical primate native to the Amazon basin, with a range extending into Central America. Based on pelage color and shape, there are two major subgroupings (Roman and Gothic) that are divided into four main species (*S. boliviensis, orstedii, sciureus,* and *ustus*) and a number of subspecies. *S. sciureus* is second to the common marmoset as the most frequently cited neotropical primate used in biomedical research. The NCRR supports a large breeding colony of squirrel monkeys at the University of South Alabama through a P40 Center grant. Animals are periodically available and imported from South America, and cost substantially less to purchase and house than macaque species. Squirrel monkeys are members of the family Cebidae, subfamily Saimiriinae, and differ from the Callitrichinae (such as the common marmoset) in many important respects, including immunology and phenotypic response to a variety of disease-causing agents. Squirrel monkey social structure is unique in utilizing seasonally based sexual segregation. Reproductively, the species lacks the twinning and chimerism commonly observed in Callitrichinae.

The squirrel monkey fills an important niche in biomedical research. The animals are tractable, easily trained, and have good manual dexterity. For these reasons, squirrel monkeys are used extensively in neurobiology programs, and models have been developed for primate-neuronal development, Prion (Creutzfeldt Jacob) disease, Alzheimer’s disease, and drug addiction. In addition, squirrel monkeys are used in infectious disease research models of human T cell lymphotropic virus type I (HTLV-I) and malarial infection in humans. These infectious diseases have a significant impact on human health. HTLV-1-associated myelopathy (or tropical spastic paraparesis) is a chronic neurologic disorder with slowly progressive and spastic lower-limb palsy. The disease is characterized by chronic progressive inflammatory changes involving predominantly the spinal cord that result in white-matter degeneration. Malaria is a leading killer of children under five and a major contributor to adult morbidity in sub-Saharan Africa; more than 300 million clinical cases and 1.2 million deaths occur each year. These diseases disproportionately affect underdeveloped regions, and an effective vaccine for both would benefit a large portion of the world’s population. This goal has been difficult to achieve, and squirrel monkey models offer unique opportunities to advance this objective.

The availability of research tools is a critical issue when evaluating species choice in model development. One of the reasons that rhesus macaques are attractive as an animal model is the availability of research tools developed by diverse groups that have application across scientific disciplines. There are clearly not as many research tools available for squirrel monkey research as there are for macaque species. The availability of a whole-genome sequence would add additional tools and speed the development of molecular and immunological assays that would positively impact diverse research programs. For vaccine research, a better understanding of MHC organization would have an immediate and positive impact.

High Priority

4) Owl monkeys (*Aotus* sp.)

Owl monkeys (*Aotus* sp.) are another neotropical primate within the family Cebidae. Until recently, owl monkeys were considered a single species (trivirgatus) with multiple subspecies. Based on coloration, karyotype, and geographic distribution, it is now thought that the genus contains a number of distinct species that may be divided into two groups: the gray-necked owl monkeys (*hershkovitzi*, *trivirgatus*, *vociferans*, and *A. lemurinus* and subspecies) that are found north of the Amazon River, and the red-necked owl monkeys (*A. miconax*, *nancymaae*, *nigriceps*, *azarae* and subspecies) that are found south of the Amazon River. Owl monkeys are nocturnally active, and have evolved changes in the eye and brain to adapt to lower-light levels. The NCRR supports a large breeding colony of owl monkeys at the University of South Alabama's Center for Neotropical Primate Research and Resources. Animals continue to be captive bred, and may be imported through a program initiated by the Pan American Health Organization.

Due to their unique nocturnal behavior and adaptive changes, owl monkeys have historically been used for primate-related vision research. Owl monkeys have been used extensively in infectious disease research, and are the most widely used nonhuman primate model of malaria. As indicated above, malaria remains one of the most important infectious diseases of mankind, and infects more then 300 million individuals annually. Owl monkeys fill a unique role, in that they are susceptible to both *P. falciparum* and *P. vivax* and are considered by many investigators as the most appropriate challenge model. As a result, they are used routinely in vaccine and therapeutic studies. They appear to be uniquely sensitive to a number of viral agents, including oncogenic gammaherpesviruses such as Epstein-Barr virus. When infected with Herpesvirus Saimiri, this species develops a malignant lymphoma resembling Burkitt's lymphoma of humans. In addition, they are highly sensitive to cytolytic effects of alphaherpesviruses, such as herpesvirus simplex 1 and herpesvirus tamarinus. For this reason, safety evaluation of simplex-based viral vectors is still required in owl monkeys by the Food and Drug Administration prior to clinical trials in humans. The genetic basis for owl monkey susceptibility to both alpha- and gammaherpes viruses is unknown. In addition to these animal models, owl monkeys suffer from a number of unique diseases that may be potentially exploited to investigate significant human health issues. These include the interrelated constellation of clinical findings including hypertrophic cardiomyopathy, hypertension, and glomerulosclerosis. Recent work in owl monkeys has suggested that the process represents a neurally based essential hypertension with involvement of the perifornical nucleus of the lateral hypothalamus.

The number of reagents available for immunologic research in owl monkeys lags behind that available in Old World primates like the macaques. While human homologies for some *Aotus* MHC, immunoglobulin, T-receptor, and cytokine genes have been published, infectious disease models would be greatly strengthened by the availability of additional genomic data.   Such information would likely strengthen ongoing vaccine studies, and provide a basis for understanding the sensitivity of the species to plasmodial and herpesvirus infections. The owl monkey sequence would complement the already available genomic sequence for *P. falciparum*.

5) Vervet or African green monkey (*Chlorocebus aethiops*)

Vervets or African green monkeys are a diverse species of African Old World primates within the subfamily Cercopithecine and genus *Chlorocebus*. Their natural range extends throughout much of sub-Saharan African, with five or six subspecies recognized by different taxonomists. They were introduced to several of the Caribbean islands in the late 1600s, where they now number in the tens of thousands. Like many other African primates, vervets are infected with indigenous strains of simian immunodeficiency virus (SIV). However, in contrast to humans infected with HIV-1 and macaques infected with various SIVs, vervets do not develop progressive loss of CD4 T cells and AIDS. As a result, research groups utilize vervets to investigate this nonpathogenic virus-host relationship, and have demonstrated a central role for activation-induced cell death in AIDS pathogenesis. In addition to this work, *C. aethiops sabaeus* is utilized extensively in behavior, endocrine, and Alzheimer's research. Major projects concerning alcohol abuse and developmental psychology have also used vervets. They are not endangered, and are bred and imported for research from two of the Caribbean islands. Several moderately sized colonies are found in the U.S.

Vervets have been promoted as an alternative nonhuman primate to rhesus macaques for biomedical research. They are considerably less expensive, do not harbor B virus, and are more tractable. Their use is hampered by a relative lack of research tools and synergy between groups. Additional sequence data would help advance a number of models, and would be particularly useful to those groups using vervets in infectious disease and AIDS-related research.

6) Chinese rhesus macaque (*Macaca mulatta lasiota and M.m. sanctijohannis*)

The rhesus macaque (*Macaca mulatta*) that has already been sequenced was a captive-born animal from a colony of pure-bred Indian-origin animals. However, as part of the whole-genome analysis, the Rhesus Macaque Sequencing Consortium surveyed genomic polymorphism in >150 kb of ENCODE sequence across nine Chinese-origin and 38 Indian-origin rhesus monkeys. The results from these analyses clearly demonstrated that there are significant genetic differences between the two populations of rhesus macaques. While there is substantial diversity within each population, only a relatively small proportion of the single-nucleotide polymorphisms are shared between the two geographic populations. The data clearly indicate that the Chinese population is genetically distinct from the Indian population, and this is entirely consistent with prior studies of mtDNA sequence, MHC polymorphism, and other genetic analyses. This Chinese versus Indian genetic difference has important consequences for biomedical research, because Indian-origin and Chinese-origin animals do not exhibit the same response to infectious diseases (e.g., SIV) or the same patterns of physiological and behavioral variation. The recent sequencing of ENCODE regions in nine Chinese-origin animals as well as previous studies of mtDNA demonstrate that the Chinese population of rhesus macaques is probably more genetically diverse than is the Indian population. Two subspecies are recognized within the Chinese region, and morphological characteristics define that difference. Hence, it is likely that biological diversity that is currently underappreciated awaits description within the broader distribution of “Chinese-origin” rhesus monkeys. It is reasonable to predict that detailed genomic data would provide evidence that one population or the other is more suitable for specific biomedical studies, as Indian-origin animals are now widely regarded as more valuable for AIDS research than are Chinese-origin animals.

Thus, given the critical importance of rhesus macaques as a model organism (rhesus macaques are the most frequently used of all nonhuman primates), it would be tremendously beneficial to have a better and more detailed understanding of the sequence differences between Indian-origin animals (now available) and Chinese-origin animals. The goal in obtaining sequence information from a Chinese-origin animal would be to generate a product of sufficient quality to identify both features that are shared and features that distinguish the two geographic forms. Ideally, one would obtain a full draft genome of one or both species. However, because costs are still high, we propose here that a different approach be pursued, that would leverage the similarity of these genomes with those of the already-sequenced *Macacca mulatta* genome. We are hesitant to be very specific because the most fruitful approach would probably involve coverage with Solexa or 454, and the appropriate depth of coverage is difficult to determine in advance of actually obtaining some pilot data. In addition, it will be useful to include some initial low coverage in the equivalent of fosmid paired-end sequencing to detect differences in gene content and structural variation between closely related species. This general “two pronged approach” is discussed further in Part 3 of this report.

Strong Justification

7) Pigtail macaque (*Macaca nemestrina*)

Pigtail macaques (*Macaca nemestrina*) are medium-sized Old World monkeys that are closely related to rhesus macaques, but not as similar to rhesus as are cynomolgus macaques. Pigtail macaques are native to Borneo, Sumatra, Malaysia, Burma, Thailand, and Vietnam. This species is available in modest numbers within the U.S., with breeding colonies operating at the Washington NPRC and the Yerkes NPRC. Animals can be imported from Asia when needed.

Pigtail macaques are most commonly used in infectious disease research, especially HIV/AIDS-related studies. They are also used for investigations in reproductive biology, endocrinology, and strategies for protecting from sexually transmitted diseases such as microbicides. Whole-genome sequencing of the pigtail macaque genome would generate novel opportunities to examine these disease models and to compare gene content and sequence between this and the other macaque genome sequences that are or will be available.

8) Sooty mangabey (*Cercocebus atys*)

Mangabeys are a group of African Old World primates within the subfamily Cercopithecine subfamily and genus *Cercocebus*. They are found throughout sub-Saharan equatorial Africa in tropical rain forests and wet lands. Like the vervets, mangabeys are infected with a number of distinct SIV strains which, despite robust viral replication, fail to cause CD4 depletion and AIDS. The sooty mangabey (*Cercocebus atys*) in particular has been used to investigate this host-pathogen relationship. Because they are endangered, many forms of research are difficult or impossible to perform on sooty mangabeys. Nevertheless, there is a breeding colony at the Yerkes National Primate Research Center, and some research groups study naturally occurring forms of the disease. Additional genomic data would likely strengthen this work and allow comparisons to the related *Chlorocebus* genus.

Two Additional Primate Species Offer Some Biomedical Value and Also Help Inform Our Understanding of Genome Evolution in the Human Lineage

We also propose two additional primate species for consideration under a biomedical rationale. Although these were not seen by the working group to be of more than moderate current priority based solely on their use as biomedical models, as described in the next section of this report they have considerable value due to their phylogenetic position in adding to our insights into annotating the human genome and improving our understanding of the evolution of the human genome. The group believed that these species should be put forward now on the combined strength of the two types of rationale. The biomedical aspects are described in this section. Given, the previously approved 2X shotgun sequences for these genomes, only an incremental 4X shotgun and appropriate BAC finishing would be required.

The biomedical importance of these species is as follows:

Bush baby (*Otolemur garnetti)* is a small nocturnal prosimian primate that emerged as a model for neurobiology and vision research (Yamada *et al*., 1999). Bush babies breed well in captivity and small colonies have been established in the United States (e.g., Vanderbilt University). They are easier and cheaper to house and maintain than their larger cousins. They are not endangered (Eichler & DeJong, 2002).

Mouse lemur (*Microcebus murinus)* provides a unique model of aging in non-human primates. *M. murinus* has been used for the study of normal brain aging and the biochemical dysfunctions occurring in age-associated neurodegeneration (Bons et al., 2006). In recent years, data have emerged suggesting that mouse lemur may be a useful model for neurodegenerative disorders such as Alzheimer’s (Gilissen *et al*., 1999) and bovine spongiform encephalopathy (Bons *et al.*, 1999). Due to its usefulness in brain research, mouse lemurs might also become an increasingly important model for the development of novel treatments and may lead to a surge in biomedical research on this species (Eichler & DeJong, 2002).

2. Bons N, Rieger F, Prudhomme D, Fisher A, Krause KH (2006). “Microcebus murinus: a useful primate model for human cerebral aging and Alzheimer's disease?”. *Genes Brain Behav.* **5**(2):120-30.
3. Bons, N., Mestre-Frances, N., Belli, P., Cathala, F., Gajdusek, D.C., and Brown, P. 1999. “Natural and experimental oral infection of nonhuman primates by bovine spongiform encephalopathy agents”. *Proc. Natl. Acad. Sci.* **96:** 4046-4051
4. Gilissen, E.P., Jacobs, R.E., and Allman, J.M. 1999. Magnetic resonance microscopy of iron in the basal forebrain cholinergic structures of the aged mouse lemur. *J. Neurol. Sci.* 168: 21-27

Part 2: Sequencing Primate Genomes to Annotate the Human Genome Sequence and to Understand Evolution of the Human Genome

Preface: This part of the overall proposal puts forward three species based on a subset of the potential rationales in this category for sequencing primates. These were species for which the working group felt that the rationales were already strong enough that they should be listed now as a priority, rather than waiting for additional analyses that could support a comprehensive exploration of species that could be proposed based on evolutionary or human comparative genomics rationales. As discussed in Part 1, two of the three species proposed here also have biomedical rationales. The working group held a wider discussion of the considerations relating to what could be learned about the human genome from additional primate sequences, some of which is presented in Appendix 3. As part of that wider discussion, members of the working group held somewhat divergent views about the significance of the justifications. However, all agreed on the three proposed species, and the desired quality of the sequence to be obtained.

The first discussion below (2a) focuses on comparative annotation of the human genome. The second (2b) very briefly discusses considerations related to the reconstruction of evolutionary events within primate lineages.

2a. Annotating the Human Genome

Background

Fewer mutations are acquired over time in functional genomic regions compared to non-functional regions. The sequence in functional regions is so important to the existence of the organism that most random changes are detrimental to survival and, thus, do not persist to become fixed in the population. These regions are said to be under negative selection (also called purifying selection), with protein-coding regions being the most familiar and most studied.

Comparison of the human- and mouse-genome sequences showed that ~5% of the human-genome sequence was under negative constraint but only ~1.7% was currently identified as functional (mostly protein coding). Further analysis suggested that 24 additional mammalian sequences were necessary to obtain the statistical power to detect all remaining regions (≥6 bp) under constraint. Such a large number of species were needed because, even though a pair of mammalian species is ~75 million years diverged, a non-functional site still has a high (~80%) chance of remaining unchanged. This gives rise to many false-positive predictions of constraint since, in reality, insufficient time has elapsed for mutations to occur. As we expand our comparison to include more and more mammalian sequences, non-functional regions acquire many more random mutations than functional ones. With 24 mammalian sequences, a point is reached where the false-positive rate is only 1 element every 10 kb (a more reasonable false-positive level).

With the aim of identifying the 5% under selection, genomic sequence data are currently being generated from 24 mammals. Although this will reveal to us genomic regions under constraint across all (or most) mammals, the question remains of how much human functionality these regions will account for. In other words, are there other regions of the human genome that give rise to functions that are not represented elsewhere in the majority of the mammalian clade? It is a reasonable assumption that a certain proportion of them will show equal constraint in closely related species (i.e., primates). Furthermore, not only will detection of these regions add to our catalog of functional bases, these regions are also highly interesting to our natural curiosity about what makes us distinctively human.

Calculations about Primate Genomes

We know that 24 mammalian-genome sequences are needed to define all mammalian-wide regions (≥6 bp) that are under constraint. Given that the statistical power is greater between sequences from mammals than primates (mammalian sequence has an 80% chance of being the same, whereas primate sequence has ~95% chance – i.e., a much greater chance of false positives), we can calculate two things :

1) What is the total available statistical power (given in total, unique branch length) available if we sample the whole primate phylogenetic tree?

2) If we assume the same false-positive rate as for the mammalian analysis (1 false positive every 10 kb), what is the minimum size of region we could hope to detect and would that be useful given what we know about existing functional regions (i.e., are they small enough to be useful)?

Box 1 shows that if species are sampled at all major branches across the primate tree, we could hope to use 25 species with a total branch length of ~1.3 substitutions/site. Adding further species helps very little, as it is the total added branch length that increases power, which, if the number extends over 20, is only a few percent of the total.

Compared to the total branch length in the mammalian low-redundancy sequencing project (~4 substitutions/site), this is a significant decrease in power. If we keep our false-positive rate the same, the minimum element size increases to ~32 bp for the available primate branch length. This is quite abit larger than the mean element size of ~15-20 bp found within the comprehensive analysis of the ENCODE regions. On the other hand, comparisons of the sequences from human, mouse, and dog genomes to that of monodelphis identified a considerable fraction of lineage specific elements (~30% of elements, ~50% of bases) over 32 bp in size. For reference, of elements present in both marsupials and eutherians 56% (87% of bases) are over 32 bp in size. Thus, analysis of 25 primates with maximum branch length could be expected to identify roughly half of the lineage specific sequence or tens of thousands of elements.

Recommendations

While the above calculations suggest that it will not be possible to find all novel primate-specific elements, it is possible that as phylogenetic shadowing or other methods looking at the increase or decrease of constraints of single bases within the mammalian elements will enable detection of important changes. **However, our current recommendation is that three species, bushbaby, mouse lemur, and tarsier be added to the list of selected primates that were proposed above based on biomedical relevance.** The rationale for adding these species is that for a relatively small sequencing investment, it will allow us to study the presence and evolution of mammalian elements through all the major nodes on the primate lineage. These genomes will also serve as outgroups if changes are found and functionally studied in later branches of the primate lineage. Furthermore, it will give the opportunity to estimate, on a genome-wide basis, the evolutionary parameters and similarities between prosimians and other primates. This information may allow us to better determine if prosimians show enough similarities to other primates to warrant the search for novel, albeit large, primate-specific elements that also reside on the prosimian branches.

2b. Reconstructing Primate Genome Evolution

The interest in reconstructing the ancestral primate genome was discussed extensively in a previous report by the AHG working group, and was met with some enthusiasm by the coordinating committee and Council for the species that were proposed at that time. Although we will wait for a future opportunity to re-introduce those arguments in detail, they should not be neglected in support of some of the species proposed in the current report. In general, high quality genome sequence from widely spaced species will aid our understanding of the evolution of the human genome. Bushbaby and mouse lemur genomes will inform the reconstruction of the ancestral genome of the Strepsirrhini. The tarsier genome will aid in understanding the ancestor to the human genome that represents the Haplorrhini. Similar arguments can be made with regard to the Owl monkey and the vervet (discussed in Part 1).

Box 1: Potential Selection of Additional Primate Genomes to Sequence

In short, the amount of sequence divergence available in the primate part of the phylogenetic tree is ~1.3 substitutions per site if 25 genomes are analyzed. This would permit detection of a 32'mer using the Eddy method.

Sampling of Major Primate Nodes

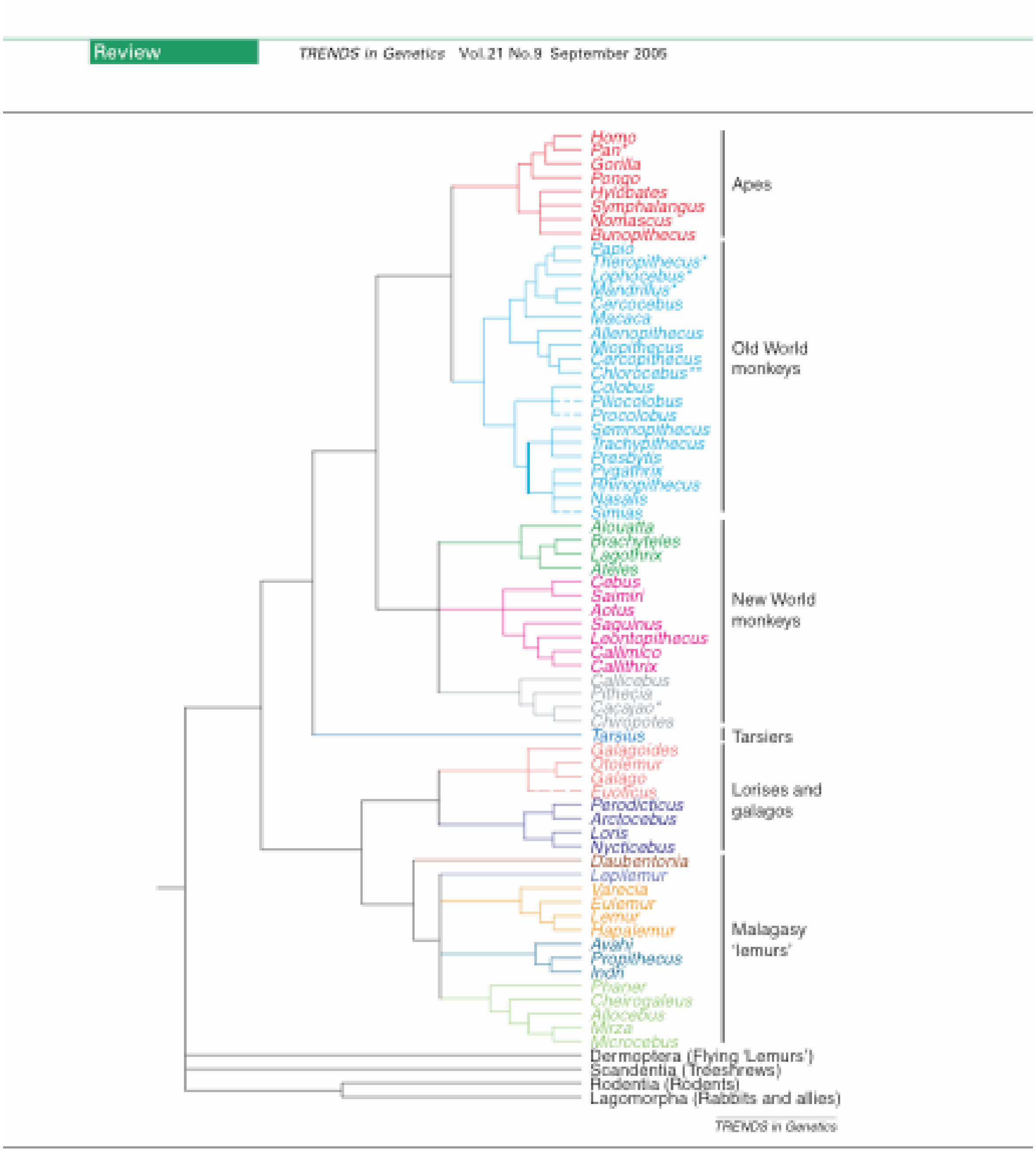
	Selective Sampling of Primate Tree			
	Deep groups	First species (sub/site added)	Additional Groups (sub/site)	Branch length (sub/site from major groups)
Lemur	4	**0.08**	0.06	0.26
Aye-aye	1	**0.06**		0.06
Loris/Bush Baby	4	**0.16**	0.06	0.34
Tarsier	1	**0.25**		0.25
NWM	6	**0.071**	0.04	0.271
OWM	5	**0.036**	0.015	0.096
Gibbon	1	**0.024**		0.024
Orang	1	**0.018**		0.018
Gorilla	1	**0.006**		0.006
Chimp	1	**0.009**		0.009
Human	1			
		Total (25 genomes):		1.33
		Already sequenced (5):		0.29
		Additional (20 genomes):		0.82

Principle of Sampling Each Node and Estimating Divergence

Branch length values come from Bill Murphy's phylogenetic tree. For each major branch, we approximated the number of deep groups based on the attached primate tree (Goodman et al. TIG Sept 2005). The estimate of additional branch length contribution per deep group was assigned by taking 2/3 of the value of the second sequenced species from that branch.

Existing genome sequences from human, chimpanzee, macaque, and marmoset comprise ~0.12 substitutions per site of sequence divergence. Additional sequencing of three hominids (orangutan, gorilla, gibbon) and representatives from deep groups in each major monkey clade (four OWMs and five NWMs) would only yield a cumulative total ~0.31 substitutions per site. Further diversity in the primate lineage is more effectively derived from the Strepsirrhini and Tarsius. Representatives from each major prosimian node would require ~10 genomes and yield 0.91 subst/site. However, the amount of prosimian genome that is informative of primate-specific adaptations may be limited.

Coverage of all significant nodes of the primate lineage would require at least 25 genomes and represent 1.33 substitutions per site, assuming deep coverage to extract all available information from each genome. Functional element discovery based on multiple alignment and statistical assessment of selection would only be able to identify elements of ~32'mer length using the Eddy method. Clearly, there are other important questions that can be answered by such a data set, but the identification of small regulatory elements will not easily be achieved using solely primate-sequence data.



Part 3: Other Justifications for Sequencing Primate Genomes

Understanding Within-Species Variation

Note: May be extendable to variation between closely-related species

There are two major reasons why one would want to characterize within-species variation in some primate species. First, if a species is a major biomedical model system (e.g., one used in pharmaceutical research), then it would be extremely valuable to understand the basis for variation in response to a treatment or susceptibility to a disease. Second, primates for which this information is available could be a good model for the population genetics of variation in humans. In addition to rationales that are of the most-direct interest to NIH, this information has the potential to provide insight into questions relating to behavioral and other phenotypic differences among primates, and to conservation biology.

How can we find such variation, and what specific insights can be gained from it? Most primates whose genomes are sequenced will be outbred individuals, although with a possibly greatly reduced population breeding size compared to wild animals. This will still lead to a substantial amount of detectable within-individual single nucleotide substitution variation (SNVs) and deletion/insertion variation (DIVs). For example, the dog genome of Tasha was 60% homozygous and 40% heterozygous, but from the 40% heterozygous fraction, 768K SNVs were observed[1]. The dog-genome sequencing effort also included additional breeds as well as other canids, allowing further insights into canid variation, such as haplotype structure within Tasha, within different dog breeds (using additional genotyping across 224 dogs and within 10 randomly selected 15-Mb regions), and inferences about ancestral haplotype structure. So as additional primate species are sequenced, detection of a significant number of within-individual SNVs and DIVs will be a windfall from the overall sequencing effort to create an assembled genomic sequence.

If a fraction of the sequencing resources were applied to very light sequencing of additional individuals or geographically separated individuals from a species, this would generate an even-more valuable variation resource. New DNA sequencing technologies (e.g., Solexa) allow very high coverage of additional individuals for variation discovery. However, such pursuits would require gaining access to more individuals, and this may be difficult in the case of species that are rare or endangered. Even if access to additional individuals for a species is limited, the variation information detected from a single individual can give some fundamental measures of variation for that species. For example, heterozygosity can be estimated for the autosomes and X (if a female is sequenced). The pattern of heterozygosity across the genome can be calculated in windows down to 100 kb in size, which can be scanned to find regions of recent bottlenecks as evidenced by decreases in heterozygosity.

Comparison of closely related species can be useful to detect regions of increased copy number or rearrangements, and some of these sites may still be polymorphic in the species being considered. For example, the depth of alignment of the chimpanzee-genome sequence reads to the human reference genome can be used to detect genomic regions that have replicated in the chimpanzee lineage relative to human. Most of these chimpanzee genomic amplifications will be fixed, but some could well be polymorphic. Fosmid- or BAC-end sequencing from one species can be aligned to another closely related species to find regions of large-scale (multi-kb) deletion/insertion differences or genomic rearrangements[2].

These considerations lead to several conclusions about how to best take advantage of the opportunities presented by these considerations. First, even without additional work, simply having good-quality draft genome sequence for outbred species will provide useful information about variation, although the data would be biased. In addition, for species where samples from perhaps 10 additional individuals can be obtained, short-read technologies can be used to add further valuable data (very low-coverage sequencing) at minimal cost. Obtaining some of this coverage (~0.3X) as fosmid end sequencing (or the equivalent) would add information about structural variation. If the species is suspected of having population structure, sampling from the known subpopulations should be done. However, to gain information about patterns of linkage disequilibrium, one would need to fully sequence targeted regions (e.g., ENOCODE) in roughly 30 individuals.

The Working Group concluded that additional reads to find variation, using this tiered approach, should be pursued with all the species that are significant biomedical model systems, where a full draft genome sequence is approved or already available.

1. Lindblad-Toh, K. et al. Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog. Nature 438, 803-819 (2005).

2. Newman, T.L. et al. A Genome-Wide Survey of Structural Variation Between Human and Chimpanzee. Genome Research 15, 1344-1356 (2005).

Appendix 1: Progress on approved primate sequencing targets (See Attached Spreadsheet)

Appendix 2. Annotating the Human Genome Working Group Primate Proposal April, 2007: Summary of Proposed Species and Priorities

Organism	Priority	Biomedical Value	Value for Annotating the Human Genome; Evolutionary/Comparative Genomics; Other	Use of the Sequence	Closely Related Sequence Available?	Proposed Type of Sequence
Baboon (*Papio hamadryas*)	Immediate Priority	Genetics of disease, Mechanisms of numerous diseases, Transplantation, Neurobiology, Reproduction	Old World major branch	Genomics of disease, primate genome evolution	Rhesus diverged about 8-9 MYA from baboon	High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Cynomolgus macaque (*Macaca fascicularis*)	Immediate Priority	Infectious disease, Pharmacology, Cancer, Diabetes, Cardiovascular disease		Genomics of disease, primate genome evolution	Cynos diverged from rhesus less than 3-4 MYA	High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTss
Squirrel monkey (*Saimiri* sp.)	Immediate Priority	Neurobiology, Infectious disease, Malaria	New World major branch	Disease models, primate genome evolution	Marmosets diverged from squirrel monkeys about 16-18 MYA	High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Owl monkey (*Aotus* sp.)	High Priority	Vision research, Other neurobiology, malaria, cardiovascular disease	New World major branch	Disease models, primate genome evolution	Owl monkeys diverged from marmosets about 16-18 MYA	High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Vervet or African green monkey (*Chlorocebus aethiops*)	High Priority	HIV/SIV, Behavior, Neurobiology, Developmental psychology	Old World major branch	Disease models, Primate genome evolution	Vervets diverged from rhesus about 9-12 MYA	High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Chinese rhesus macaque (*Macaca mulatta lasiota and M.m. sanctijohannis*)	High Priority	Many disease models, including diabetes, cardiovascular, osteoporosis, neurobiology		Disease models, primate genome evolution	There is a small but biomedically significant genetic divergence between Indian and Chinese origin species	Coverage w/ new seq tech sufficient to compare with *M. mulatta;* light FES coverage or equivalent to identify structural variants
Pigtail macaque (*Macaca nemestrina*)	Strong Justification	Infectious disease		Disease models, Primate genome evolution	Pigtail macaques diverged from rhesus less than 4-6 MYA	High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Sooty mangabey (*Cercocebus atys*)	Strong Justification	Infectious disease (HIV/SIV)		Disease models, Primate genome evolution	Mangabeys diverged from rhesus about 8-9 MYA	H High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Tarsier *(Tarsier synrichta)*	High priority		Major node; prosimian or New World	Primate genome evolution		High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Bushbaby (*Otolemur garnetti)*	High priority	Neurobiology and vision research	Major prosimian node	Disease model, primate genome evolution		High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs
Mouse lemur *(Microcebus murinus)*	High priority	Neurodegenerative disease, biological rhythm, cerebral aging	Major prosimian node	Disease model, primate genome evolution		High-quality draft coverage (~6-7X), Refinement of selected BACs, ESTs

Appendix 3: Sequencing Primate Genomes to Annotate the Human Genome Sequence: An Overview of Important Considerations

One important reason for sequencing primate genomes is to improve our understanding of the human genome sequence by identifying genomic intervals that are functionally important. In theory, these intervals can be detected because they evolve over evolutionary time in a manner that can be distinguished from neutral (non-functional) DNA. Although sequence data are already being generated from numerous mammals precisely for this purpose, primate-genome data are needed in cases where the function (or the mechanism by which the human genome attained that function) arose in an ancestral primate. In such cases, comparisons with a non-primate genome will not reveal the signature of non-neutral evolution. These regions are of particular interest because they are where the human genome has gained its unique capabilities.

Genome comparisons between humans and non-primate mammals typically look for regions conferring a function that is common to all mammals, or at least to the mammals being studied. These regions are presumed to be under *negative selection* (also called *purifying selection*). That is, many or most random mutations in the interval decrease reproductive fitness, and are thus removed from the population. As a result, the interval changes more slowly over evolutionary time than do non-functional regions. With short functional intervals (e.g. <30 bp) or intervals where the functional constraint is weak, sequence data from several mammals are required before the region can be reliably identified. There is a well-developed theory for how many species are needed to reach a given level of resolution, which puts a premium on total branch length (i.e., on picking species that are well distributed over the phylogenetic tree) and preferring fast-evolving species (see details below). This same theory can be applied to intervals that came under negative selection in primates. In summary, detecting primate-specific negative selection (by a process sometimes called *phylogenetic shadowing*) favors obtaining sequence data from several dozen primates.

A novel and exciting facet of primate-sequence data is that it can lead to identification of genomic intervals that have recently come under *positive selection* (also *called adaptive evolution*) in the lineage leading to humans (i.e., where some mutations have increased reproductive fitness by improving the genomic interval's functional capability). In particular, there is considerable interest in learning the genetic basis that accounts for features that are unique to humans, which by definition will not reflect functions shared with all primates; this goal has focused attention on finding regions that show signs of positive selection by comparisons between humans and other primates, particular chimpanzee (with another primate sequence used to tell if a human-chimpanzee difference occurred in the lineage leading to humans). The ideal qualities of sequence data for finding regions under recent positive selection are quite different from those for identifying negative selection. A hallmark of positive selection is a rate of sequence change that exceeds that of neutral DNA, so occasional regions of low sequence quality can masquerade as the regions being sought— a situation that would confound the comparative analyses. Moreover, genomic regions that have recently undergone large-scale duplications or rearrangements (i.e., precisely the ones that are difficult to assembly using whole-genome shotgun sequencing methods) are particularly likely to contain intervals under positive selection. Thus, the goal of identifying and analyzing fast-changing regions of the human genome supports the internal NHGRI report dated January 2006 and entitled *Non-human Primate Genome Sequencing*, which "shares with previous recommendations the general aim of pursuing fewer index primate species sequenced at higher quality, rather than many at low quality."

At this time, it is difficult to quantify how much will be learned about functional regions in the human genome from extensive primate-genome sequencing. It has been estimated that at least 5% of the human genome has been under negative selection throughout mammalian evolution. However, we know of no comparable figure for the fraction under primate-specific negative selection or the fraction under positive selection in recent human ancestry. ***Such unknown features are of profound importance for developing a sophisticated understanding of human genome evolution and function, and it is now clear that deeper, high-quality primate-genome sequencing will be central to acquiring that understanding.***

Principles of Detecting Sequences Under Selection

Most methods designed to identify sequences under (negative or positive) selection are based on the general principle of characterizing the signatures of neutral drift in some way, and identifying statistically significant departures from these signatures. Methods for identifying evolutionarily conserved elements, dN/dS-based tests for positive selection, and the McDonald-Kreitman test all follow this general approach. Note that the method of Pollard et al. (2006) is similar, but uses a null model representing uniform negative selection rather than neutral evolution. The differences among methods come down to what signatures are used (e.g., substitutions, indels, SNPs), what proxy is used for neutrally evolving bases (e.g., synonymous sites, ancestral interspersed repeats), how is the statistical test constructed (e.g., likelihood ratio test, test based on counts of events), and what are the candidate elements under consideration (conserved elements, genes, sliding window, all possible intervals via an HMM). In all cases, the effectiveness of these methods in detecting selection depends on certain general questions:

1. How good an indicator for selection are the signatures that are used? To put it another way, how likely is it that a sequence under selection will show no detectable signature, or that a sequence not under selection will have a spurious signature?

2. How well do the candidate elements match real functional elements? That

is, are most real elements considered as candidates (with approximately the right boundaries), and are a substantial fraction of candidate elements real (of concern because of multiple hypothesis testing)? This is an especially important question with regard to noncoding sequences. A related issue is how many bases are actually under selection within each candidate element.

3. How good is the null model? The wrong null model (e.g., a model of neutral evolution that does not allow for rate variation across sites) might lead to many false-positive predictions.

Steady progress is being made on issues (2) and (3) in terms of both methods (better models, better algorithms) and data (high-throughput experimental methods such as ChIP-chip and transcription tiling arrays producing new candidates). It is issue (1) that is most relevant in selecting species for sequencing. Sequencing of additional species, or of additional individuals within a species, reveals new signatures that can help to distinguish sequences under selection from those that are evolving neutrally.

From a statistical point of view, the critical question relating to issue (1) is how to maximize *power*. In other words, given a type of signature (e.g., substitutions) and a set of candidate regions (e.g., transcribed regions, pan-mammalian conserved elements), what choice of species/individuals for sequencing will allow the most functional elements to be detected at some fixed, acceptable false-discovery rate. Power depends strongly on the phylogeny relating the species, especially (but not exclusively) its total branch length; on the lengths of the sequences under selection (or, more precisely, on the number of selected bases within them); and on the strength of selection. It also depends on assembly and alignment quality, and it depends in more subtle ways on properties such as base composition and mutation-rate variation. Increasing branch length tends to increase power, but only up to a point. In general, power is greater if branch length is distributed across many branches instead of concentrated in one or two very long branches – that is, if the tree is "bushy" rather than "willowy."

In the typical formulation, the goal is to maximize power for elements that are shared across all species. At inter-mammalian distances, power is generally maximized by reaching to the farthest possible corners of the mammalian phylogeny, within the limits of alignability. That is, total branch length is a reasonable proxy for power. This is probably true for positive selection as well as negative selection, although very strong positive selection may prohibit alignment of distant mammalian genomes.

Maximizing branch length is not necessarily the right goal, however, when considering lineage-specific selection and evolutionary turnover of functional elements. In this scenario, a functional element has undergone some change in selective pressure on some branch of the phylogeny (note that complex scenarios involving multiple branches are also possible). In order to detect this change, one needs both sufficient sequence data from descendant species (in the subtree below the branch in question) to characterize the new selective mode and sufficient sequence data from outgroup species (outside the subtree) to characterize the old selective mode. Similar principles to those discussed above hold in both the subtree and the "supertree": a bushy tree with as much branch length as possible is ideal. If the change is recent, however, as in cases of primate-specific selection, the main challenge is in obtaining sufficient information for the subtree. In addition, high-quality sequences and alignments are especially important, since many types of errors will show up as apparent changes in selective pressure.

An additional challenge that arises once changes in selective pressure have been identified is distinguishing between positive selection and relaxation of constraint. If some species show significantly faster evolution than a locally calibrated neutral rate (e.g., based on nearby ancestral interspersed repeats or synonymous sites), then positive selection is a good possibility. Levels of within-species polymorphism can also be used to identify a likely selective sweep. In addition, more specific types of evidence for adaptation might be identified, such as:

1. Radical changes that nevertheless appear to maintain some kind of function, as deduced from what we know about how the functional element works. Examples are changes that produce a new version of a protein-coding gene that still has a healthy ORF, or changes in a structural RNA gene that still produce a healthy structure.

2. Evidence of episodic positive selection, or positive selection followed by purifying selection.

3. Evidence that subsequent compensatory changes were made to improve the function of the element or to restore function that was lost. A maladaptive change that was rescued by a compensatory change is a particularly interesting change. Common examples include a frame-shifting indel in a protein followed by a frame-restoring indel, or a pair of compensatory changes in an RNA structure

Analysis of Rapidly Changing Regions of the Human Genome

With the sequencing of chimpanzee genome, considerations of positive selection have been on the upswing within the mammalian genomics community. The chimpanzee sequence is on average so similar to human (just over a 1% difference in terms of nucleotide substitutions) that there is essentially no statistical power to find modest-sized regions changing at below the neutral rate (negative selection), and moreover the main interest is in what makes us different from chimpanzees. However, the hunt for intervals under positive selection has been hampered by the low quality of the initial chimpanzee-sequence assembly and the lack of data from a close outgroup species to distinguish changes in human from changes in chimpanzee. Additional chimpanzee-sequence data that improved the overall assembly and the availability of the rhesus-genome sequence have improved the situation, and the orangutan-genome sequence should improve things even more by providing a better outgroup species. Note that gorilla/human divergence may have occurred so near in time to chimpanzee /human divergence as to limit the value of gorilla as an outgroup for human- chimpanzee comparisons; some human genomic regions are more gorilla-like than chimpanzee -like (see Ruvolo 1997).

Most efforts to identify regions under positive selection have focused on protein-coding sequences, for which the traditional method is to compute dN/dS [the ratio of non-synonymous (i.e., amino-acid changing) differences (dN) between orthologous sequences to synonymous changes (dS)]. Positive selection is indicated for genes where dN/dS is >1. Several genome-wide studies of this type comparing human and chimpanzee sequences have recently been published. The study performed for the rhesus-genome sequence is of particular relevance for this report because it addressed the relative value of using primates at different evolutionary distances from humans, as well as assessing the effect of data quality. One conclusion was that inclusion of the macaque-genome sequence substantially improves statistical power to detect positive selection in primates, compared with previous scans with just the human- and chimpanzee-genome sequences. A second conclusion was that the finished genome sequences of macaque and chimpanzee would allow the number of genes in high-confidence orthologous gene trios (a prerequisite for reliably determining if a gene is under selection) to be increased by at least 23%.

In what follows, we outline several additional approaches for using primate-genome sequence data to identify fast-evolving regions in the human genome and/or human-specific functional elements. Where data exist, we give published ‘lower bounds’ on how much of the human genome can be annotated by the approach.

One class of methods that check for selection acting on genes utilizes both between-species divergence (d) and within-species polymorphisms (p). For instance, the McDonald-Kreitman test compares the ratio of divergence to polymorphisms (d/p) at non-synonymous sites with the ratio at synonymous sites using a chi-square test for statistical significance. If d/p is higher for non-synonymous sites than for synonymous sites, positive selection is indicated. Bustamante et al. (2005) applied this test to over 11,000 human genes (with polymorphisms from three human populations compared to divergence from chimpanzee), and concluded that 9% of the potentially informative loci displayed a significant signal for positive selection, while 13.5% had a significant signal for negative selection.

A strategy for finding human-specific positive selection in non-coding DNA is to search among intervals that show high conservation among non-human mammals for cases where the sequence of human is significantly different from that of the other species [Pollard et al. 2006) report 202 genomic elements of this type, while Prabhakar et al. (2006) identify 992 such cases, though at a higher false-positive rate).

A novel use of primate-sequence data for helping to understand the human genome is explored in the rhesus-genome sequencing paper. The basic idea is that even the most quickly changing regions of the human genome (e.g., regions of variable copy number in the human population and/or where current assembly methods are inadequate to finish the sequence) may be quiescent in certain primates. Refining the sequence of the primate-genome region, which may be relatively undemanding, can reveal the structure of the ancestral region and illuminate the evolutionary path recently taken along the human lineage.

Proposals to sequence the Neanderthal genome (Green et al. 2006; Noonan et al. 2006) raise an exciting scenario for interspecies comparisons that might illuminate very recent human evolution. However, there are serious concerns about post-mortem DNA damage in ancient specimens (not to mention potential contamination from modern human DNA, economic consideration caused by the low fraction of Neanderthal DNA in the sample, and issues about destroying precious samples) that need to be better understood before this project can be endorsed whole-heartedly. Gilbert et al. (2007) showed that in a mammoth sample that remained frozen for its 28,000-year history and was then collected, approximately 4 in 1000 bases were changed by DNA damage. According to theoretical models, the rate of DNA damage increases exponentially with temperature, but empirical studies are needed to measure the actual effects of damage in a somewhat older sample (note that Neanderthals went extinct more than 28,000 years ago) from a substantially warmer climate.

Bustamante, C. D., et al. (2005) Natural selection in protein-coding genes in the human genome. *Nature* **437**, 1153-1157.

Gilbert, M. T., et al. (2007) Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Res* **15**, 1-10.

Green, R. E., et al. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330-336.

Noonan, J. P., et al. (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314, 1113-1118.

Pollard, C. S., et al. (2006) Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* **2**, e168.

Prabhakar, S., et al. (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786.

Ruvolo, M. (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol.* **14**, 248-265.

Appendix 4: Annotating the Human Genome Working Group

Evan E. Eichler	University of Washington
Richard A. Gibbs	Baylor College of Medicine
Eric Green (chair)	National Human Genome Research Institute (NHGRI)
David Haussler	University of California, Santa Cruz
Eric Lander	The Broad Institute
Steven McKnight	University of Texas Southwestern Medical Center at Dallas
Stephen O'Brien	National Cancer Institute
Maynard Olson	University of Washington
Brian Raney	University of California, Santa Cruz
Jane Rogers	The Sanger Institute
Robert Strausberg	J. Craig Venter Institute
Robert Waterston	University of Washington

Ad hoc members

Adam Siepel	Cornell University
Jeff Rogers	Southwest Foundation for Biomedical Research
Keith Mansfield	Harvard University
Svante Paabo	Max Planck Institute, Germany
Dario Boffelli	Lawrence Berkeley National Laboratory
Webb Miller	Penn State Unviersity
James Sikela	University of Colorado at Denver and Health Sciences Center