A MODIFIED VERSION OF THE PROPOSAL FROM THE WORKING GROUP ON ANNOTATING THE HUMAN GENOME

Table of Contents

I. Executive Summary of the Scientific Plan to Annotate the Human Genome

With the completion of a finished sequence of the human genome, attention is turning to the crucial challenge of annotating the genome – identifying the functional elements encoded in the genome, the distinctive forces that shaped the human genome and the relationship between genomic variation and susceptibility to disease. One of the most powerful systematic approaches for annotating the human genome is comparison with additional genomic sequence information.

The Working Group on Annotation of the Human Genome was charged by National Human Genome Research Institute (NHGRI) with developing a scientific program to produce genome sequence information to propel efforts to annotate the human genome. The Working Group's initial proposal was considered by the Coordinating Committee (CC) on Selection of Sequencing Targets, which subsequently worked jointly with the Working Group to produce this modified proposal that was then approved by the National Advisory Council for Human Genome Research (NACHGR). This document, written primarily by the Working Group, describes the modified proposal and its scientific rationale.

The proposal seeks to exploit the prodigious genome sequencing capacity of NHGRI's currently funded sequencing centers. As sequencing costs continue to drop, the capacity of the centers will continue to increase. At present, the combined capacity of the NHGRI-funded centers is ~140 Gb of raw sequence information per year. The capacity for producing finished genomic sequence is much more modest, about 0.5-1.0 Gb per year.

The proposal for producing genomic sequence to annotate the human genome consists of a three-part scientific program, which builds on the genome sequences that have been completed or are in progress at the NHGRI centers and elsewhere. Specifically, these include mouse, zebrafish, chicken, dog, opossum, platypus, macaque and chimpanzee. These species represent key branch points in the evolutionary tree of vertebrates and most are key experimental systems for biomedical research. The initial Working Group proposal strongly recommended obtaining high quality sequence for each of these animals to provide a solid foundation for the comparative analyses to follow. Specifically the group suggested such high quality sequence might be obtained by (i) deep, whole genome shotgun coverage (minimum of 6-fold) for each of these genomes, and (ii) subsequent directed efforts to close gaps, increase quality, order and orient contigs, and resolve recent duplications into separate copies. This recommendation was not included in the modified proposal, but remains for discussion by the NHGRI Council.

Component 1: Systematic identification of functional elements in the human genome through comparison with diverse eutherian mammalian genomes.

Background: The systematic identification of all functional elements encoded in the human genome (i.e., the human genome 'parts list') is a pressing need and should be one of the highest priorities for genome analysis. Based on comparison of the human genome with that of mouse and now rat, it has been possible to infer that ~5% of the human genome is under purifying selection and therefore likely to be functional; these

sequences include those for protein-coding genes, non-coding RNAs, regulatory elements and structural elements. Defining these elements with precision, however, will require comparison with many additional mammalian species. Such comparisons should allow complete definition of exons and increasingly precise definition of signals required for gene regulation and chromosome behavior.

Proposed plan: The modified proposal encompasses obtaining low-coverage genomic sequences for 15 diverse eutherian mammals. Eight of those should be sequenced to 2-fold depth now. Pending analysis of those results and additional exploration of the mammalian tree, seven more species will be chosen and sequenced to a depth based on the analysis. Coverage for these additional species might be able to be significantly less (as little as 0.5-fold) thereby minimizing sequencing capacity required. Alternatively, higher coverage of all species, perhaps 3-fold, may be required to obtain optimal annotation. The initial set of mammals proposed was chosen to optimize total evolutionary branch length in order to best identify functional sequences in the human genome. In addition the medical relevance of the cat was taken into account. Similar criteria will be used to choose the second set of organisms to be sequenced. The recommended species will sample the mammalian tree broadly, opening opportunities for a range of studies and providing important data that will guide the selection of additional genomes for which high-quality reference sequence should be generated in later years.

Choice of Eutherians.

The following were chosen for the initial 2-fold sequencing effort:

- armadillo
- guinea pig*
- European common shrew
- elephant
- hedgehog (European or African)**
- cat
- rabbit
- tenrec

(* As an alternative, the bat would contribute a shorter branch length than guinea pig, but potentially has a much smaller genome (2.1 vs. 3.8 Mb?) and samples an important lineage; it may thus be somewhat more efficient. However, specific bats need to be investigated for branch length and genome size before substituting bat for guinea pig. ** The European hedgehog has been characterized with respect to branch length. If the African hedgehog can be shown to be of the same clade and branch length, it would be preferable due to better availability. Note: Improved techniques for estimating genome size would be desirable in all cases.)

Capacity required: 60-135Gb.

Component 2: Understanding distinctive aspects of the human lineage, through sequencing of primates.

Background: An important question of general interest and specific medical relevance is: "What makes us human?" As a complement to Component 1, comparative analysis can be used to define those distinctive sequences that arose along the human

lineage. These include sequences that distinguish primates from other mammals, which will allow inferences to be drawn about the events that have led to the distinctive human phenotype. The available draft sequence of the chimpanzee genome is providing initial insights, but it is clear that substantially more information is required to gain a comprehensive picture. This includes a high-quality sequence of the genomes of several primates in order to provide a comprehensive view of the significant differences in the human genome.

Proposed plan: The modified proposal recommended obtaining high quality genome sequence for one new primate, the orangutan, as an initial goal. This will bring the total number of non-human primate genome sequences expected in the next two years to four, since sequencing of the chimpanzee and macaque genomes are currently underway in NHGRI-supported centers and it is anticipated that sequencing the gorilla genome will be initiated elsewhere. These genome sequences will provide the key resources for interpreting human-chimpanzee differences. In the longer term, genome sequences from marmoset, gibbon and mouse lemur should be added to sample additional key branch points on the primate tree.

Capacity required: 40 Gb . (The precise capacity required to produce high-quality genome sequence (beyond deep-coverage draft, but short of fully finished) is uncertain. Making reasonable guesses, we estimate that it would translate to ~40 Gb for orangutan. Increased capacity will be required in later years for additional primate species.

Component 3: Obtain a systematic catalogue of human variation through re-sequencing of many additional human genomes.

Background: Identifying the genetic factors that predispose to disease is a central goal of human genome analysis. Various techniques have been developed for correlating regions of the human genome with disease (including, most recently, association studies in populations using haplotypes), but they still require tedious and time-consuming studies to identify the common variation. To accelerate the identification of disease genes and to understand our origins, a comprehensive catalog of human variation can be used for follow-up of regional studies and for direct association studies.

Recent analyses demonstrate that it would be possible to identify the vast majority (95%) of all human variants having allele frequency ≥ 1% by obtaining ~100-fold additional coverage of the human genome sequence from a diverse collection of individuals.

Proposed plan. The modified proposal recognizes the need for continued resequencing of the human genome to study human variation and its effect on disease and endorses a human resequencing component. Because this is a rapidly evolving field of high significance, the NHGRI will hold a workshop to explore the various strategic and technological approaches before implementing the human resequencing component of the NHGRI program.

Capacity required: to be determined.

The Working Group, CC and NACHGR all recognized that there are uncertainties within each of the three components and recommended that the overall program and each of the components be monitored over time and that the details be adjusted in light of experience gained. To this end, the NHGRI should immediately establish a task force to

address the theoretical issues (including strategy and power) involved in extracting information from comparative genome sequence analysis. While a subcommittee of the Working Group performed some of this kind of analysis, the issues need to be revisited with more extensive modeling on existing data sets (mouse, rat and dog) than there has been time for to date and in light of the new sequence data that will continue to be generated by the program. In particular, the trade-off between quantity and quality of genomic sequence must be watched closely in order to ensure that the quality of the product is sufficient to meet the goals. In addition, the scientific returns of the programs must be evaluated on an ongoing basis to assess the value of the incremental information provided by each new genome.

The modified proposal presented above would bring us well along the path toward defining those sequences in the human genome that are conserved within most mammals, identifying those segments that have changed in the human genome relative to the rest of the primate lineage, and recognizing those genomic sites that are variable in human populations. A resource would be created that would allow investigators looking at a stretch of human sequence to know if that sequence were present in the primordial mammalian genome, if it were under selection, if it had changed significantly in the primate lineage, and if it were variable in the human population. Providing investigators with this information will be critical as we attempt to integrate the vast amount of information in the human genome into an understanding of human biological complexity and to apply this knowledge to improving human health and well-being.

II. Detailed Proposal to Identify Conserved Regions of the Human Genome (Component 1): Annotating Highly Conserved Regions of the Human Genome by Low-Redundancy Sequencing of Multiple Eutherian Mammals

A. Background and Theoretical Analysis

The identification of bases under selection requires the study of enough animals whose genomes are sufficiently diverged so that non-conserved bases are rarely the same by chance but, at the same time, are close enough to allow reliable alignment across conserved bases. Single pair-wise comparisons fail to provide the discrimination necessary to recognize all or even most conserved regions of complex genomes. For example, in human-mouse sequence comparisons, the reliable alignment of neutrally evolving orthologous sequences is challenging, yet the distribution of the inferred 5% of sequence under selection overlaps substantially with that of neutrally evolving sequences (Mouse Genome Sequencing Consortium 2002). Fortunately, the discriminatory power of sequence comparisons can be enhanced by the use of multiple genomes separated by appropriate evolutionary distances. This concept has been illustrated vividly on a genome-wide scale by recent papers on comparative sequence analyses of multiple yeasts (Cliften et al. 2003) (Kellis et al. 2003) and for selected regions from vertebrate species (Thomas et al. 2003). The objective of the proposal to sequence multiple mammalian genomes is to identify most of the conserved features/regions of the human genome.

Background Concepts

An important concept for comparative genome analysis is that of **branch length**, which provides a measure of the likelihood that a neutrally evolving base at a given site remains the same by chance. For example, the mouse-human branch length is ~0.50 substitutions per site, so that a correctly aligned base has an approximately $e^{-0.5}$ (or 60%) probability of being unchanged just by chance in the two genomes. When analyzing multiple species, the branch lengths can be combined (taking care to count each segment of the multiple branches only once) to provide an overall estimate that a base at a given site will remain unchanged by chance. If the total branch length is brought to 4 substitutions per site, a correctly aligned base would have a probability of $e^{-4.0}$ (or 2%) of being unchanged just by chance.

Signal-to-noise ratio (SNR) is another important concept. Mammalian genomes, with only ~5% of bases under selection, have about a 15-fold lower SNR than yeast, with ~75% of bases under selection. For example, one of the above-mentioned yeast studies used four species that provided a maximum pair-wise branch distance of 0.54 and a total branch length of 0.83 to extract meaningful information about regulatory sequences. An increase in total branch length from x to x+a results in improvement in the SNR of e^{a}-fold. Thus for studies of mammals to match the power of the yeast study an increase of ~3 in the total branch length would be required, improving the SNR by ~20-fold.

Reliable alignment is yet another critical concept. Larger features such as exons, with their associated evolutionary constraints on indels and rearrangements, can be

aligned readily across substantial evolutionary differences. For example, Thomas et al. (2003) demonstrated that chicken-human sequence alignments (associated with a greater total branch length than that of human-mammal) could be used to detect most exons. But reliable alignment of smaller, less constrained features (e.g., regulatory elements) depends on the alignment of flanking neutrally evolving sequences. Again from Thomas et al. (2003) and further illustrated by Margulies et al. (2003), sequences conserved across multiple eutherian mammals are not detected by chicken-human comparisons, even with high-quality sequence that provides confident alignments anchored by conserved exons. The effectiveness of sequences from non-eutherian mammals, such as a marsupial (e.g., opossum; providing a branch length of 0.78) and a monotreme (e.g., platypus; providing an additional branch length of 0.44), at identifying conserved non-coding regions is currently being evaluated (by the NIH Intramural Sequencing Center and the laboratory of E. Green). While these latter sequences will almost certainly prove valuable for finding the most highly conserved mammalian genomic elements, their greater evolutionary distance and distinct phenotypic differences will limit their utility for identifying all functional genomic elements common to eutherian mammals.

Sequence Considerations

High-quality finished sequence clearly provides the most complete information for performing genome comparisons. However, the cost of producing sequence at the quality standards established for the human genome is still substantial and finishing capacity is limited in the NHGRI-supported and other sequencing centers. For example, the mouse whole-genome sequence assembly used ~20B raw bases (~8-fold redundancy), but despite ongoing efforts for the past 20 months at three sequencing centers, the available finished sequence is now only approaching 1.5 Gb.

The above-cited yeast studies suggest an alternative strategy. While Kellis et al. (2003) used a deep-shotgun strategy to produce highly contiguous sequence, Cliften et al. (2003) used only 2- to 3-fold redundancy, aligning the individual sequence reads directly to the finished *S. cerevesiae* sequence. Importantly, the ability of the two approaches to define sequences under selection was comparable. Together, these results suggest that attaining lower-redundancy coverage (and therefore lower-quality sequence) of a larger number of genomes might prove more effective for finding highly conserved genomic regions than using the same amount of resources to generate a smaller number of high-quality sequences.

Indeed, preliminary work by several groups suggested that individual mouse sequence reads can be usefully aligned to the human genome sequence. Since the mouse-human branch length is among the greatest of those between humans and other eutherian mammals, sequence reads from other mammalian species should align at least as reliably. Assemblies derived with even 2-fold shotgun coverage should yield somewhat greater continuity and thus improved pair-wise alignments. Thus, obtaining low-redundancy sequence (with no additional finishing) from an adequate number of eutherian mammals should represent a viable path forward for gaining the additional total branch length needed to detect highly conserved genomic elements in the eutherian mammalian genome.

To assess the utility of low-redundancy sequence coverage of multiple mammalian genomes in identifying highly conserved genomic regions, the working group formed a sub-group involving individuals from several institutions. A manuscript describing their analyses is being prepared for publication. The key conclusion from these studies is that 2-fold sequence coverage from multiple additional mammals would be highly effective at detecting highly conserved genomic regions. Specifically, their two main areas of study and conclusions were:

- **Aligning sequence from low-redundancy data sets.** For low-redundancy sequence coverage from multiple species, the worst-case scenario will be to have individual reads of 600-700 bases with which to work. Based on simulated and actual sequence data, the sub-group's analyses indicated that alignment of the sequence data of the planned depth to the finished human sequence will generally be sufficient to detect conserved regions of interest. Furthermore, with 2-fold sequence coverage, 73% of a given species' genome will be assembled into fragments >600 bases, with 36% of the genome residing in supercontigs >42,000 bases.

- **Detecting highly conserved sequences using low-redundancy data sets.** Actual data sets of 2-fold sequence coverage from multiple species are able to detect ~90% of the highly conserved genomic regions [Multi-Species Conserved Sequences or MCSs; see Margulies et al., 2003] at 97% specificity, assuming some small number of sequences are available in high-quality draft forms (e.g., dog, mouse, rat, and chicken). The improvements encountered with even slightly higher levels of sequence coverage were fairly minor.

While further study is required to understand all of the relevant issues and trade-offs associated with the low-coverage, multi-species strategy, these studies provide evidence that data sets of low-redundancy shotgun coverage from multiple eutherian mammals would be valuable for cataloging those regions of the human genome that are conserved among most mammals. At the same time, these initial studies illustrate the complexity of analyzing large collections of sequence from evolutionarily diverse species. Indeed, the generation of such data sets will provide the raw material for fueling this important and rapidly growing area of bioinformatics.

The Working Group, Coordinating Committee and NACHGR are aware that this proposal runs counter to the sequencing projects supported by NHGRI in the past, specifically those aiming to generate the complete sequence of 'reference' genomes, in particular those of model organisms. The latter provide foundational information for large research communities, and each imperfection in a sequence leads to a price paid by the user community. Similarly, the NHGRI has supported the generation of high-quality genome sequences of species at important branch points on the evolutionary tree. The

need for high-quality sequence for additional critical branch-point species such as chicken, opossum, and platypus must also be considered.

However, most mammals without sequenced genomes are studied by much smaller groups, whose numbers pale in comparison to the number of investigators studying humans and the major experimental mammals. Nonetheless, the former would be greatly empowered by the availability of low-redundancy shotgun sequence of their genome(s) of interest. The resulting data would allow the sequences of most exons to be identified, greatly facilitating follow-up studies. Including BAC- or fosmid-end sequences would frequently allow retrieval of the corresponding clones, providing experimental access to any genomic region of interest. The many additional genome sequences produced through the proposed strategy would be available for computational and experimental analyses, inevitably empowering a broader group of investigators to pursue genomics-based studies in previously under-studied species.

B. Proposal

Previous and ongoing sequencing of eutherian mammalian species will together provide ~45-fold coverage (measured in increments of ~3-Gb genomes), providing an evolutionary tree with total branch length of ~1.0 substitutions per base (see Table 1). Platypus and opossum genome sequences will add ~1.2 additional branch length, but it is unclear how effectively these genomes will reveal sequences conserved among eutherian mammals. The modified proposal involves generating an additional ~20-30-fold coverage or more from eutherian mammals (to provide ~60-95 Gb of total sequence). The first set of mammals alone will increase the total branch length to 3.65 (3.84 substitutions per base within eutherian mammals).

Specifically, the modified proposal involves:

(1) Selecting a set of eight eutherian mammals, which will increase the total branch length by ~1.5 substitutions per base. The specific choices for the first set are given in Table 1 and Figure 1 below.

(2) Generating ~2-fold sequence coverage of each of the genomes of the mammals in the first set, with rigorous ongoing evaluation of the data. This will involve an amount of sequencing roughly equivalent to generating deep (i.e. ~8-fold) coverage of 2 mammals, but will provide much greater total branch length.

(3) Generating more complete sequence for each selected species across the 44 ENCODE regions. This will provide higher-quality data sets for ~1% of the mammalian genome that will be used for more in-depth evaluation of this approach. Furthermore, this would allow the sequencing effort to effectively dovetail with the broader ENCODE Project, which aims to establish rigorously the best route towards finding all functional elements in the human genome by first focusing on a selected 1%. Several strategies for generating high-quality sequence across the ENCODE regions for each species can be envisioned; at a minimum, the same BAC-based targeted approach already being employed by the ENCODE Project can reliably be used.

(4) Evaluating the results from steps 1-3 and, as warranted, repeat with a second set of 7 species at a depth of coverage from 0.5-2 fold. Alternatively, deeper coverage could be obtained from the species in both sets, as deemed appropriate.

Choice of Species

The specific species in the first set were chosen primarily for their ability, in combination, to annotate the human genome. The criteria included:

- Phylogenetic distance from human **and** other sequenced organisms, e.g. mouse.
- Maximization of total substitutions/site (branch length) across the tree while still keeping in mind the value of the particular organism/clade sampled to the broader community outside of annotation alone (i.e. what other significant biological questions could be informed by having a 2X sequence: see other).
- Availability of sample (e.g. are the species particularly elusive, requiring substantial efforts to obtain).
- The candidate species' overall morphology and how this can be used to help understand developmental mechanisms (i.e. sequencing from developmentally extreme placental mammals).

Other considerations included:

- Biomedical relevance to human biology
- Economic importance, such as agricultural species
- Ancestral or primitive genome organization
- Genome size
- Research community and background studies to date

Similar criteria should be applied in choosing the second set of mammals.

Figure 1. Partial mammalian tree with different species sets highlighted as follows: black – human; blue – sequenced or in progress; red – proposed set 1; green – other mammals; pink – non-placental mammals.

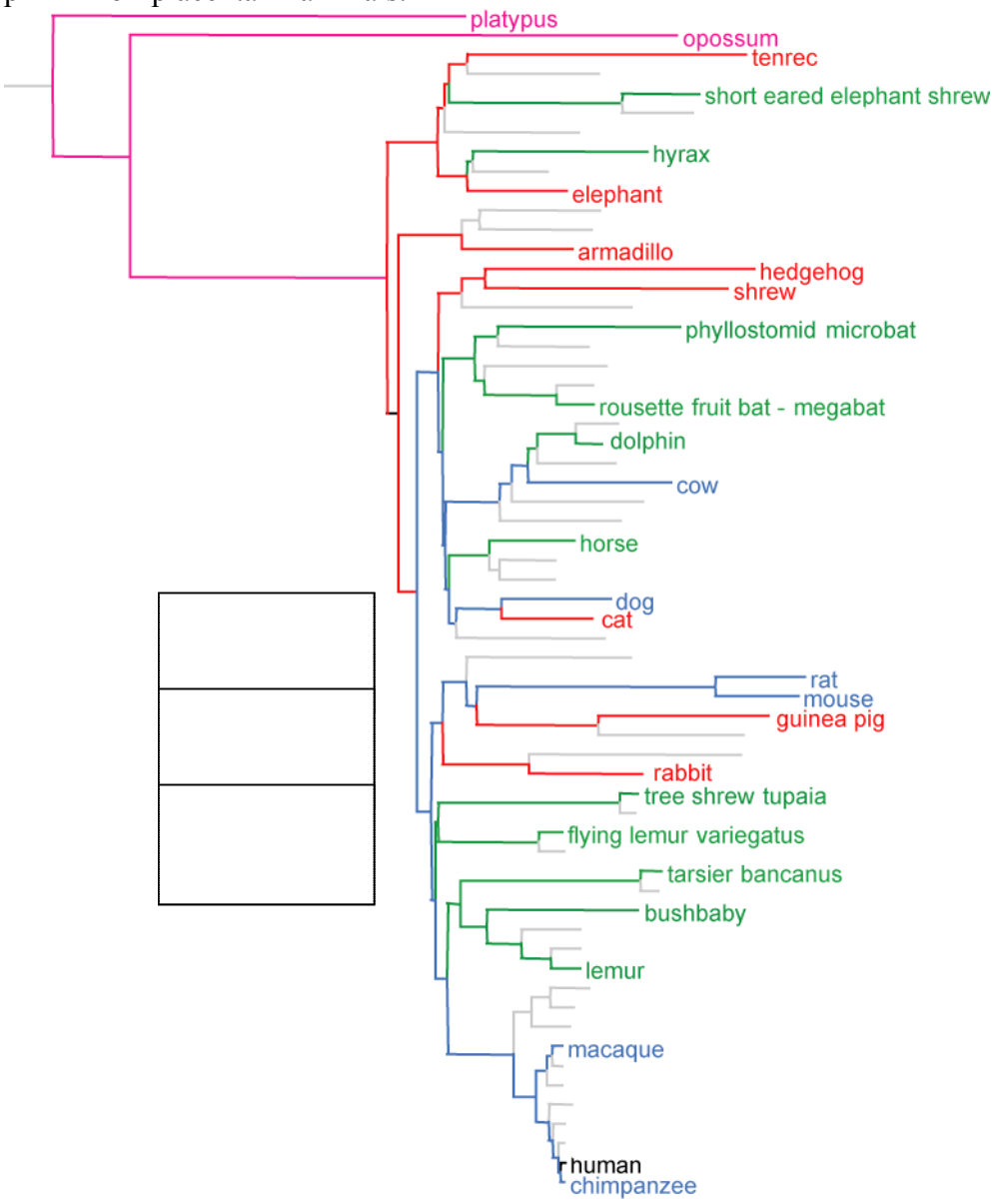


Table 1. Mammalian Species

Sequenced/in progress genome	**I:Dist. to ancestral placental mammal**	**II:Dist. to ancestral node of superordinal clade**	**III: Unique branch length**
Human, *Homo sapiens*	0.16	0.12	0.16
Mouse, *Mus musculus*	0.36	0.32	0.32
Rat, *Rattus norvegicus*	0.35	0.31	0.08
Dog, *Canis familiaris*	0.20	0.15	0.19
Chimp, *Pan troglodytes*	0.16	0.12	0.01
Cow, *Bos taurus*	0.26	0.21	0.20
Macaque, *Macaca mulatta*	0.16	0.12	0.03
Opossum, *Monodelphis domestica*	0.75	NA	0.75
Platypus, *Ornithorhyncus anatinus*	0.44	NA	0.44
Total branch length			**2.18**
Group 1 First priority			
African savannah elephant, *Loxodonta africana*	0.17	0.13	0.17
hedgehog tenrec, *Echinops telfairi*	0.32	0.28	0.27
Nine-banded armadillo, *Dasypus novemcinctus*	0.15	0.09	0.17
Lagomorph: rabbit	.024	0.20	0.19
Domestic cat, *Felis catus*	0.18	0.13	0.08
European hedgehog, *Erinaceus europeaus*	0.33	0.28	0.28
Guinea pig, *Cavia porcellus)*	0.34	0.30	0.26
European common shrew, *Sorex araneus)*	0.31	0.26	0.22
Total branch length			**1.64**
Total overall branch length			**3.82**

Notes: Included are two Tables of mammalian species: 1.) Species already identified for high quality whole genome sequence (>6-fold); 2.) Species selected as first priority for 2-fold sequencing for human genome annotation based upon a consensus of the working group and coordinating committee. The primary criterion for selection was the branch length contributed by the species, i.e. sequence divergence each selected species achieves from human and in parallel from the common ancestor of all placental mammals (26 orders including 18 placental orders, 7 marsupial orders and one monotreme order). For context, we included the estimated mean genome sequence divergence rate (as number of substitutions per base pair for each selected species) based upon the large mammal phylogeny data set (1-3). Three columns include : I.) **Distance to ancestral placental mammal,** i.e. to the common ancestor of the eighteen orders of placental mammals; II.) **Distance to ancestral**

node of superordinal clade, i.e. to the common ancestor of the super-ordinal clade/lineage to which each species belongs (Afrotheria, Xenarthra, Euarchontoglires, and Laurasiatheria); III. Distance of added species to the already existing lineage coverage of previously selected species. Relatively low values for some species (e.g. chimp, macaque, rat, etc.) reflect that a lineage divergence represented by the genome sequence of previously selected closely related species (human, human and mouse respectively) was not included in the additional lineage length contributed by the new species. Additional criteria for species selection are discussed in the text and in reference 4.

1.) Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S.J.: Molecular phylogenetics and the origins of placental mammals. Nature. 409: 614-618, 2001.

2.) Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C., Teeling, E., Ryder, O.A., Stanhope, M., de Jong, W. W., and Springer, M. S.: Resolution of the early placental mammal phylogeny using Bayesian phylogenetics. Science 294: 2348-2351, 2001.

3.) Springer, M.S., Murphy, W. J. Eizirik, E., O'Brien, S.J.: Placental mammal diversification and the K/T Boundary. Proc. Nat. Acad. Sci. USA 100:1056-1061, 2003.

4.) O'Brien, S.J., Eizirik, E. and Murphy, W.J.: On choosing mammalian genomes for sequencing. Science. 292:2264-2266, 2001.

GROUP 1

1. **African savannah elephant, *Loxodonta africana.***

   One of the four major placental lineages, Afrotheria was the first to diverge from the other major groups of placental mammals. Thus, from a phylogenetic standpoint, afrotherians represent a unique group of modern placental mammals whose origins date back to the deepest node of Placentalia. Elephants represent the most highly visible and best-studied species of any afrotherian, in fields ranging from ecology to behavior and genetics. We recommend obtaining sequence from the African savanna elephant (*Loxodonta africana*), rather than the Asian elephant because of the greater presence of African elephant research and data in the PubMed and GenBANK databases. Also, the savannah elephant is remarkably monomorphic relative to the Asian and Forest elephant species; this will be useful in assembling the elephant genome sequence. A morphologically derived placental mammal (i.e. extremely modified from the ancestral mammalian body plan), elephants exhibit large body size, long life span, and complex behavior. An elephant sequence would be useful for understanding genomic alterations affecting body growth and macroevolutionary developmental changes. A BAC library is currently in production (VMRC-11) from a male individual (McClean) from the Disney Wild Animal Park.

2. **Lesser hedgehog tenrec, *Echinops telfairi***

   Tenrecs represent another divergent member of the superorder Afrotheria, and a unique order of endemic African insectivores (Afrosoricida). The tenrec contrasts with elephants by exhibiting a number of ancestral mammalian morphological features, including low & variable body temperature, a cloaca, and undescended testicles. For annotation purposes, the tenrecs examined thus far have among the most accelerated nucleotide substitution rates of all placental mammals.

3. **Nine-banded armadillo, *Dasypus novemcinctus.***

   The armadillo provides a representative of the second major placental clade to diverge, Xenarthra. The nine-banded armadillo was selected as an initial xenarthran representative because it is the best known and studied of all armadillos, and perhaps of all xenarthrans, in the fields of physiology, genetics, and ecology. In addition, the NIAID has previously established breeding colonies of nine-

banded armadillos for leprosy research. Furthermore, this species displays a number of interesting reproductive characteristics, including single-sex quadruplets and delayed embryo implantation, which would provide interesting models for comparative genetic analysis. A BAC library (VMRC-5) has been constructed from a wild-caught male.

4. **Rabbit, *Oryctolagus cuniculus***

As a representative of the order Lagomorpha, the sister group to rodents, the rabbita is in widespread use as an experimental/laboratory mammal; thus, a 2X genome sequence would generally benefit the biomedical research community. Rabbits are well studied for immune system gene organization/function, for a remarkable myxoma virus-based eradication program in Australia, and in toxicology research.

5. Domestic cat, *Felis catus*.

The domestic cat provides the second single most useful biomedical research model within the mammalian super-order Laurasiatheria, after the dog. Its extremely conserved genome organization will provide insight into the maintenance of long stretches of conserved synteny over evolutionary time. The human genome is relatively conserved like the cat, while the dog genome has been shuffled fourfold relative to the carnivore ancestor and most other mammals. The human and cat genomes have the most highly conserved gene order of placental mammals. Well-developed cat genomic resources (linkage maps, radiation hybrid map, full MHC sequence, BAC library RPC-86) in conjunction with a 2X sequence would allow rapid identification and characterization of genes that contribute to conditions for which the cat is an important animal model, including over 200 human hereditary diseases, infectious diseases (e.g. Feline immunodeficiency virus, Feline leukemia virus, SARS-like coronavirus), and reproductive biology suitable for stem cell, transgenics and knockouts.

6. **European hedgehog, *Erinaceus europeaus***

A second representative of the order Eulipotyphla, hedgehogs have the benefit of one of the most accelerated nucleotide substitution rates outside of rodents.

7. Guinea pig, *Cavia porcellus*.

The guinea pig provides sampling from the caviomorph lineage within Rodentia. In addition to its accelerated nucleotide substitution rate, the guinea pig is widely used in studies of physiology, toxicology, infectious disease, cardiovascular disease, diabetes, etc. For these purposes, a genome sequence will be extremely useful for experimental research . The highly accelerated nucleotide substitution rate of this species (similar to mouse and rat) led to the erroneous conclusion that guinea pigs were not rodents (Graur et al. Nature 1991) .

8. European common shrew, *Sorex araneus*.

Another mammal with an accelerated DNA substitution rate, a shrew sequence would provide representation of the ancestral mammalian morphology within the major clade Laurasiatheria. The European common shrew is the most advanced eulipotyphlan characterized at the genomic level. In addition, it exhibits an interesting feature among mammals, in that chromosome number varies considerably across individuals of the same population. The presence of XYY males would provide a unique genomic perspective on the evolution of the Y chromosome in mammals. A BAC library for this species is in production at Clemson University Genome Institute/Arizona Genome Institute.

References

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen, and M. Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301: 71-6.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241-54.

Margulies, E.H., M. Blanchette, D. Haussler, and E.D. Green. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* 13: 2507-18.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-62.

O'Brien, S.J., E. Eizirik, and W.J. Murphy. 2001. Genomics. On choosing mammalian genomes for sequencing. *Science* 292: 2264-6.

Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell, B. Maskeri, N.F. Hansen, M.S. Schwartz, R.J. Weber, W.J. Kent, D. Karolchik, T.C. Bruen, R. Bevan, D.J. Cutler, S. Schwartz, L. Elnitski, J.R. Idol, A.B. Prasad, S.Q. Lee-Lin, V.V. Maduro, T.J. Summers, M.E. Portnoy, N.L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C.P. Brinkley, S.Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S.L. Ho, M.C. Huang, E. Karlins, P.L. Laric, R. Legaspi, M.J. Lim, Q.L. Maduro, C.A. Masiello, S.D. Mastrian, J.C. McCloskey, R. Pearson, S. Stantripop, E.E. Tiongson, J.T. Tran, C. Tsurgeon, J.L. Vogt, M.A. Walker, K.D. Wetherby, L.S. Wiggins, A.C. Young, L.H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C.L. Shu, P.J. De Jong, C.E. Lawrence, A.F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E.D. Green. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424: 788-93.

III. Detailed Proposal to Identify Human-specific sequences in the Human Genome (Component 2): Annotating the Human Genome using multiple Primate Genome Sequences

Comparison of the human genome to those of other primates can reveal the changes that have occurred in fashioning the human genome from the primordial mammalian genome, and in particular from our last common ancestor with chimpanzee. Our anthropocentrism, of course, makes us curious about what distinguishes us from other primates. More practically, human-primate comparisons will have implications for understanding human disease (Olson and Varki, 2002). The comparisons should also allow us to begin to address a central question in evolutionary biology: what types of changes in what types of genes account for the emergence of a new species so conspicuously an outlier among great apes?

Unlike most comparative genomics to this point, human-primate comparisons focus on what is different rather that what is conserved. Accordingly, errors, gaps in the sequence and failure to sort out segmental duplications confound the analysis as they masquerade as differences. Consequently, most human-primate comparisons are best carried out with high-quality, comprehensive sequence. High-quality sequencing of primate genomes will provide a balanced view of genome variation (including regions of structural variation, segmental duplication, lineage-specific events and chromosomal variation) as a function of evolutionary time.

A. Background

.

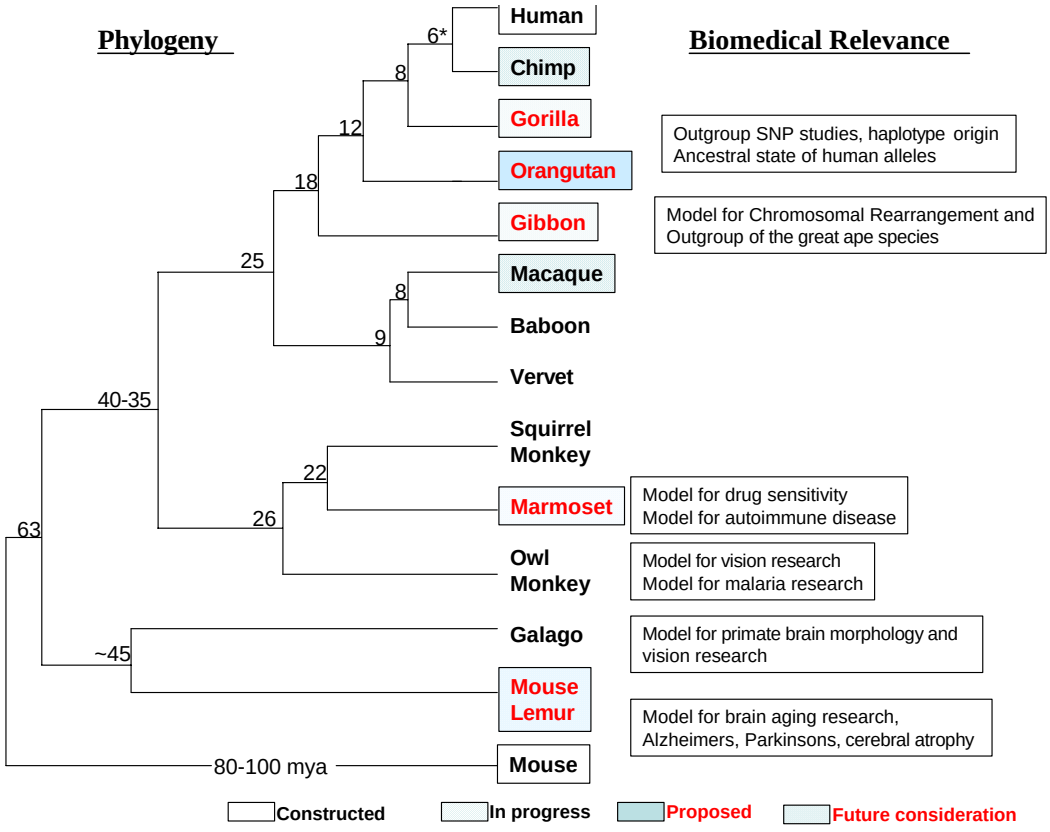
The comparison of the draft chimp sequence with the human reference sequence reveals many intriguing differences, including rapidly evolving genes, gene loss and differences in interspersed element activity. It provides as well a view of some of the mutational forces at work in our genomes. But the draft nature of the chimp sequence creates important limitations for these analyses. Although the long-range continuity is high and coverage of the chimp genome approaches 95%, the number of gaps in the sequence remains large. Almost 50% of genes are missing one or more exons. Other genes contain frame-shift differences, the majority of which appear to be errors.

Gene annotation is only one aspect of genome annotation. Rates, patterns and mechanisms of genome variation (segmental duplication, gene conversion, etc.) are others of important biomedical relevance. But sequence and assembly artifacts associated with the 4-fold redundant assembly complicate the identification of subtle rearrangements (gene deletions, inversions, new insertions, areas of positive selection). Recent segmental duplications can only be assayed indirectly and the relatively rare rearrangements between the two genomes are difficult to distinguish from possible misassemblies.

In addition, with only rodent genomes available as outgroups for these two closely related genomes, it is often difficult to establish the ancestral state of a region.

Even for genes, Clark et al. (2003) were only able to establish 1:1:1 orthology (mouse:chimp:human) for 7,645/~20,000 genes and the consortium currently analyzing the publicly available chimp genome sequence has experienced similar limits. Furthermore, at most positions, one can only conclude that the two genomes differ, without being able to decipher which lineage has changed.. Differences in rates of change are even more difficult to establish. A set of high quality primate genomes is needed to address these issues unambiguously.

Primate Phylogeny. With respect to human, most agree that groups of non-human primates diverged at six distinct evolutionary timepoints (6, 8, 14, 18, 23-25, 35-40, 55-60 million years ago), corresponding to chimpanzee, gorilla, orangutan, hylobatids, Old World monkeys, New World Monkeys and prosimians. (Figure 1)



Specific Uses of Primate Genomic Sequence:

- **SNP annotation**— Reconstructing ancestral state of single nucleotide polymorphisms and human haplotypes. Recently published SNP studies emphasize the value of genomic sequence from non-human primates to determine the ancestral and derived status of human alleles (Chen and Li 2001; Kaessmann et al. 2001). Data from species (chimp, gorilla and orangutan) closely related to human are particularly valuable in eliminating ambiguity with respect to the ancestral status of a common human polymorphism. Sequence from these species provides a critical backdrop for testing the impact of genetic drift and rapid expansion on the frequency and structure of contemporary human haplotypes. Such data will also allow the identification of ancient alleles that now occur as minor variants within the human population. These analyses require at least one outgroup (gorilla or orangutan) in addition to the chimpanzee and human sequences; a second outgroup would reduce ambiguities further. Disease and evolutionary mutations of functional significance can be related —i.e. pyrin and familial mediterannean fever (Schaner, 2001), FOXP2 and autism. (Enard, 2002).
- **Gene annotation**— Establishment of orthologous relationships among multigene families over long evolutionary distances is virtually impossible (Dehal, 2001). Only ~50% of genes can be mapped to mutual-best 1:1:1 mapping positions between human, chimp and mouse (Clark et al., 2003). High quality sequence from another great ape and a more distant primate (macaque) will dramatically improve identification of orthologs. In addition, these sequences will allow a direct assessment of pseudogenization (Paabo, Gilman, 2003). Determination of which genes have been gained and lost during the course of human evolution is of critical biomedical and evolutionary relevance (Olson and Varki, 2002). Great-ape sequences will establish factors (gene conversion, duplication, rearrangement) that can confound 1:1 mapping of genes.
- **Segmental Duplication Structure and Variation**. ~4% of the human genome exists as blocks of duplication >95% that are >20 kb in length. Data suggest a burst of activity over the last 25 million years, probably since the divergence of new and old world monkeys. (Bailey, 2002; Eichler, unpublished). Segmental duplications contribute to disease phenotypes (25 microdeletion/microduplication syndromes) (Stankiewiecz, 2003). They are also preferential sites of chromosomal breakpoints (30-50%) between humans and mice (Armengol, 2003; Bailey 2004) and between humans and apes (50-70%; Eichler and Rocchi, unpublished). Transcript density is enriched within these regions (280,000 of the 2.2 million best spliced ESTs) (Eichler and Furey, unpublished). Comparison of

the human genome with those of the great apes and with selected, more distantly related primate genomes will shed light on the evolution of these complex structures.

- **Patterns of Non-neutral Selection** (adaptive selection and ancestral polymorphisms) a) Adaptive--A complete set of genes from humans, two hominoids and an outgroup (human, chimp, orangutan, macaque) will allow the adaptive evolution of all genes to be interrogated. b) Non-human primate sequence will pinpoint genomic regions that appear potentially polymorphic within both species (KIR/HLA loci). Among immunologists, for example, comparative sequencing between human and non-human primates has been used to provide compelling evidence for models of balancing selection regarding genes associated with human blood group antigens (Grimsley et al. 1998; O'Huigin et al. 2000).
- **Mechanism of Chromosomal Evolution**: The tempo of karyotype evolution (defined here as the number of syntenic rearrangements per unit time) is 10X faster among hylobatids. What is the underlying genomic mechanism? The relatedness of hylobatid genome sequence (4% divergence) and the high quality of the human genome reference provides the ability to interrogate the junctions precisely and provide insight into the molecular events underlying karyotype evolution.

B. Full Proposal

In addition to the finished human sequence, the planned high-quality chimpanzee sequence and the anticipated high quality draft sequence of macaque, at least one more great ape genome should be sequenced. The working group proposed that the NHGRI-supported centers should sequence the orangutan genome because gene trees are less likely to conflict with the species tree.** Furthermore, the gorilla genome will likely be sequenced elsewhere. Recognizing the high cost of finished sequence and the limited capacity of NHGRI centers, a lesser quality standard is acceptable for the orangutan genome, in which >98% of the sequence is represented, gaps are no more frequent than 1/100kb and segmental duplications are accurately assembled. One approach to achieve this would be an initial 6X whole genome shotgun (WGS) data set (18 Gb per genome), followed by automated closure strategies. In addition, the whole genome assembly would be complemented by BAC-based finished sequence within WGS-intractable regions (segmental duplications, large-scale deletions, inversions, insertions, subtelomeric transition regions, pericentromeric transition regions). For hominoid genomes (based on a current assessment of the chimpanzee genome), this might require ~2,000 BACs, spanning 350Mb of sequence. The WGS data should include paired-end sequence data from 8-fold redundant (clone coverage) BAC libraries, as well as additional fosmid clones. Large insert mate-pairs are required to span areas of segmental duplication and provide more genomic coverage to confirm sites of large-scale inversion, deletion and insertion. BAC libraries have already been constructed for the orangutan.

**Both the gorilla and orangutan are important, immediate outgroups for the human-chimp differences. Chen and Li (2001) have argued that a high fraction of gorilla/chimp/human gene trees (~30%) disagree with the species tree, presumably because the last coalescent for the genes existed prior to the gorilla/human/chimp split. Regardless of the exact fraction, their work points out limitations of using only a single closely related species as an outgroup.

The Working Group concluded that, for an outgroup, genomic sequence from either gorilla or orangutan would be sufficient for comparisons with the human genome, and that the availability of both sequences would be very useful. The principal use of an outgroup genome sequence would be to determine the ancestral state of single nucleotide polymorphisms. The Working Group is aware that at least one other sequencing group may be considering the gorilla genome as a target and encourages that group and its funding source to take on the project. Consequently, the Working Group recommended that NHGRI focus on the orangutan (Pongo pygmaeus) sequence.

Orangutan. In addition to the reasons given above, there are several aspects of the orangutan sequence that make it an attractive target for the NHGRI sequencing program. Its estimated divergence from the human lineage (12-14 mya) places it at an evolutionary midpoint between human and Old World monkeys (separation 25 mya) (Chen and Li 2001; Goodman 1999). It is, therefore, considerably sought after for comparative sequencing for molecular evolutionary analysis and for testing for models of selection. Among immunologists, for example, comparative sequencing between human and orangutan has been used to provide compelling evidence for models of balancing selection regarding genes associated with human blood group antigens (Adams et al. 1999; Bontrop et al. 1991; Otting et al. 1998).

The orangutan karyotype is the best representative of the ancestral hominoid state. Both human and African ape chromosomes are believed to be largely derivative, requiring a minimum of 10-15 chromosomal rearrangements from this hominoid archetype (Muller and Wienberg 2001; Yunis and Prakash 1982)

The majority of intrachromosomal duplications (>80% by bp representation) are > 98% identical . Most of the available Human Genome Project data suggest that the bulk of duplications occurred after the separation of the orangutan but before the trichotimization of the African apes. Targeted analysis of these regions in orangutan has been used to reconstruct the ancestral origin of several segmental duplications and to infer the series of events that have created this duplication architecture in humans (Eichler et al. 1996; Jackson et al. 1999; Johnson et al. 2001; Monfouilloux et al. 1998; Orti et al. 1998; Zimonjic et al. 1997)

Two subspecies of orangutan are recognized: Bornean (*P.p.pygmaeus*) and Sumatran (*P.p.abelli*). Genetic data from both subspecies suggest extensive polymorphism (Warren et al. 2001; Zhang and Ryder 2001). Unlike human and most African great

apes, there is no evidence for a recent genetic bottleneck in the population history of this species. Coalescent ages of 1.1 –2.1 million years have been proposed for orangutan alleles (nearly 10-20 fold that of human), providing a critical backdrop for testing the impact of genetic drift and rapid expansion on the frequency and structure of contemporary human haplotypes.

In the longer term, there are compelling reasons to obtain genome sequence for the following additional primates.

1. Gorilla (*Gorilla gorilla*) (expected to be sequenced elsewhere)
***Relevance*:** The gorilla is now recognized as an outgroup species to human and chimpanzee since it diverged 1-2 mya prior to the separation of these sister taxa (Goodman et al. 1999). The principal use of a gorilla genome sequence would be to determine the ancestral state of single nucleotide polymorphisms. The phylogenetic proximity virtually eliminates back mutation and parsimony determination of ancestral state. This is particularly relevant in regions of unusual selection, i.e. HLA antigen loci, where comparative sequencing has been used to resolve the evolution of immune-related genes. Most molecular evolution studies require a third organism to root trees that include the human and chimpanzee comparison (Kaessmann et al. 2001; Mathews et al. 2001). Long-range PCR amplification of genomic regions has proven difficult in this regard with many amplicons >10 kb in length (~10%) failing to PCR. Finally, areas of rapid evolutionary turnover (subtelomeric, pericentromeric, and large-low copy repeats) have changed radically over short periods of time.
Source: Three subspecies of gorilla are recognized (Western Lowland, Eastern Lowland and Mountain Gorilla). Western Lowland gorillas are the most common and least endangered, and they were selected for the contruction of a BAC library (CHORI255).

2. Gibbon (*Hylobates*).
Relevance: The gibbon represents a phylogenetic link between the great apes and the Old World monkeys. It provides a unique view of genomic temporal change between 15-20 mya of species separation (human and gibbon). This organism demonstrates an accelerated rate of karyotype evolution--compared to other primate and most mammals (Muller et al. 1997; Muller and Wienberg 2001). Comparative studies indicate an unusually large number (n>45) of chromosomal rearrangements when compared to hominoid species. Unlike most hominoids, these karyotypes have been subjected to a large number of fission events. Comparative sequencing would be used to understand the molecular basis for chromosomal rearrangements—i.e. the transition region and sequences that may have predisposed to such events. Detection and sequence characterization of such large-scale rearrangements requires an abundance of paired-end sequence to satisfactorily traverse regions enriched in common and low-copy repeat sequences. Information obtained from such studies could provide valuable insight into both germline and somatic chromosomal instability associated with chromosomal rearrangement.
Source: There are at least five different species or subspecies belonging to the genus Hylobates. *All show rapid evolutionary rearrangement with respect to human.* Material from *Hylobates* has recently been identified for the purpose of gibbon BAC library (Dr. Alan Mootnick at the Santa Barbara Zoo, CA, Director of Gibbon Conservation).

3. Marmoset (*Callithrix jacchus*). Alternate: Squirrel Monkey (*Saimiri sciureus)*
***Rationale*:** This organism is a member of the New World Monkeys (Superfamily Ceboidea), estimated to have diverged from the anthropoid common ancestor (35-40 mya). It is an anchor species of the callitrichine clade, one of seven anciently separated New World monkey clades that diverged from each other at least 18 mya (Chiu et al. 1996). This species is a key organism for studies related to immunity, drug sensitivity and brain function. Its small size, fecundity and inexpensive handling make it one of the non-human primate models of choice. This species is commonly used to assess the toxicological effects of various drugs and has, on occasion, been shown to be a more appropriate model than rodents in which to test adverse drug reaction or long-term side effects (Carey et al. 1992; Jackh et al. 1984). Immunological

studies have shown that the marmoset immune system is a particularly good model when compared to other primates for testing antibody specificity and recognition. Marmosets have been used to develop models of multiple sclerosis, an autoimmune disease of the central nervous system (Genain and Hauser 2001; Hart et al. 2000), as well as autoimmune colitis and thyroiditis.
Source: The most commonly used marmoset subspecies in research is *Callithrix jacchus jacchus*. The animal is not endangered. Several large colonies exist within the United States including 235 animals at the Wisconsin Regional Primate Center and ~80 animals located at the Southwest Regional Primate Center.

4. Malagasy gray mouse lemur (*Microcebus murinus*). Alternate: (*Lemur catta).*
***Rationale*:** The primate order may be divided into two major divisions: prosimians and anthropoids. Ancestral prosimians diverged ~60 mya from the primate lineage leading to the ancestors of New World monkeys, Old World monkeys, apes, and humans. Despite a massive extinction of prosimian species in the late Eocene (50 mya), remarkable diversity still exists (43 species are currently identified). Although several species have acquired adaptive specializations to specific ecological niches, prosimian features are generally regarded as more primitive. The prosimians occupy a unique position both morphologically and phylogenetically in the primate lineage (Goodman et al. 1998). Evolutionarily, they are regarded as the outgroup of all simian species and the link to more "primitive" mammalian orders (Insectivora and Chiroptera). From the perspective of molecular evolution studies, a very strong argument can be made for genome sequence from this group. There are at least 2 major divisions of prosimians (galago and lemur). *Microcebus murinus* is biomedically representative of the latter. Over the last 10 years, this species has emerged as a model for aging research. The organisms are small (50-80 g), short-lived, fecund (2-3 offspring per year) and reach sexual maturity at a young age (10 months). *Microcebus* shows stereotypical signs of aging such as susceptibility to blindness due to lens opacity, increased frequency of tumour formation, stereotypic geriatric behavioural changes and brain lesions similar to those associated with Alzheimer's disease (Austad 1997). Histological examination of mouse lemur brains has identified the accumulation of A beta deposits within the blood vessel walls of the cortical parenchyma similar those observed in human Alzheimer patients (Gilissen et al. 1999). Several molecular studies have been initiated to recover genes associated with AD and brain aging (Bons et al. 1995; Calenda et al. 1998). Interestingly, the life span of mouse lemurs is dependent on the number of annual photocycles that the animal experiences. The average life span is 5 annual photocycles. If the photocycle is accelerated to 8 months in duration, the mouse lemur still lives only 5 cycles on average. These observations suggest that they will become important models for other studies related to the molecular mechanisms of aging (Perret 1997). Finally, in recent years, data has emerged that suggest this species may serve as a useful model for bovine spongiform encephalopathy infection (Bons et al. 1999). Its fecundity, propinquity to humans and usefulness in brain aging research has lead to a dramatic surge in biomedical research on this species. Large research centers with 200-300 individuals are maintained particularly in Europe (Brunoy and Paris, France).
***Source*:** Duke University Primate Center is "an international center for research on living and fossil primates". They have the largest and most diversified collection of lemurs in the U.S which includes a small cohort of mouse lemurs. In addition to this source, Dr. John Allman (Caltech University) has a small colony of *Microcebus.* Materials from euthanized animals are sporadically available.

References.

Adams EJ, Thomson G, Parham P (1999) Evidence for an HLA-C-like locus in the orangutan Pongo pygmaeus. Immunogenetics 49:865-71.

Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. Hum Mol Genet 12:2201-8

Austad SN (1997) Small Nonhuman Primates as Potential Models of Human Aging. Ilar J 38:142-147

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE (2004) Hotspots of chromosomal evolution. Genome Biol 5:R23

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003-7

Bons N, Jallageas V, Silhol S, Mestre-Frances N, Petter A, Delacourte A (1995) Immunocytochemical characterization of Tau proteins during cerebral aging of the lemurian primate Microcebus murinus. C R Acad Sci III 318:77-83

Bons N, Mestre-Frances N, Belli P, Cathala F, Gajdusek DC, Brown P (1999) Natural and experimental oral infection of nonhuman primates by bovine spongiform encephalopathy agents. Proc Natl Acad Sci U S A 96:4046-51

Bontrop RE, Broos LA, Otting N, Jonker MJ (1991) Polymorphism of C4 and CYP21 genes in various primate species. Tissue Antigens 37:145-51.

Calenda A, Mestre-Frances N, Czech C, Pradier L, Petter A, Perret M, Bons N, Bellis M (1998) Cloning of the presenilin 2 cDNA and its distribution in brain of the primate, Microcebus murinus: coexpression with betaAPP and Tau proteins. Neurobiol Dis 5:323-33

Carey GJ, Costall B, Domeney AM, Jones DN, Naylor RJ (1992) Behavioural effects of anxiogenic agents in the common marmoset. Pharmacol Biochem Behav 42:143-53

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444-56.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-3

Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, Land M, Terry A, Zhou CLE, Rash S, Zhang Q, Gordon L, Kim J, Elkin C, Pollard MJ, Richardson P, Rokhsar D, Uberb acher E, Hawkins T, Branscomb E, Stubbs L (2001) Human chromosome 19 and related regions in mouse: conservative and lineage specific evolution. Science 293:104-111

Eichler EE, Lu F, Shen Y, Antonacci R, Jurecic V, Doggett NA, Moyzis RK, Baldini A, Gibbs RA, Nelson DL (1996) Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. Hum Molec Genet 5:899-912

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869-72

Genain CP, Hauser SL (2001) Experimental allergic encephalomyelitis in the New World monkey Callithrix jacchus. Immunol Rev 183:159-72

Gilad Y, Man O, Paabo S, Lancet D (2003) Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A 100:3324-7

Gilissen EP, Jacobs RE, Allman JM (1999) Magnetic resonance microscopy of iron in the basal forebrain cholinergic structures of the aged mouse lemur. J Neurol Sci 168:21-7

Goodman M (1999) The genomic record of Humankind's evolutionary roots. Am J Hum Genet 64:31-9

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585-98

Grimsley C, Mather KA, Ober C (1998) HLA-H: a pseudogene with increased variation due to balancing selection at neighboring loci. Mol Biol Evol 15:1581-8

Jackh R, Rhodes C, Grasso P, Carter JT (1984) Genotoxicity studies on di-(2-ethylhexyl) phthalate and adipate and toxicity studies on di-(2-ethylhexyl) phthalate in the rat and marmoset. Food Chem Toxicol 22:151-5

Jackson MS, Rocchi M, Thompson G, Hearn T, Crosier M, Guy J, Kirk D, Mulligan L, Ricco A, Piccininni S, Marzella R, Viggiano L, Archidiacono N (1999) Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications, and unstable sequences with homologies to telomeric and other centromeric locations. Hum Mol Genet 8:205-215

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE (2001) Positive selection of a gene family during the emergence of humans and African apes. Nature 413:514-9.

Kaessmann H, Wiebe V, Paabo S (1999) Extensive nuclear DNA sequence diversity among chimpanzees. Science 286:1159-62.

Kaessmann H, Wiebe V, Weiss G, Paabo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27:155-6.

Mathews DJ, Kashuk C, Brightwell G, Eichler EE, Chakravarti A (2001) Sequence variation within the fragile X locus. Genome Res 11:1382-91

Monfouilloux S, Avet-Loiseau H, Amarger V, Balazs I, Pourcel C, Vergnaud G (1998) Recent human-specific spreading of a subtelomeric domain [In Process Citation]. Genomics 51:165-76

Muller S, Hollatz M, Wienberg J (2003) Chromosomal phylogeny and evolution of gibbons (Hylobatidae). Hum Genet 113:493-501

Muller S, O'Brien PC, Ferguson-Smith MA, Wienberg J (1997) A novel source of highly specific chromosome painting probes for human karyotype analysis derived from primate homologues. Hum Genet 101:149-53

Muller S, Wienberg J (2001) "Bar-coding" primate chromosomes: molecular cytogenetic screening for the ancestral hominoid karyotype. Hum Genet 109:85-94.

O'Huigin C, Satta Y, Hausmann A, Dawkins RL, Klein J (2000) The implications of intergenic polymorphism for major histocompatibility complex evolution. Genetics 156:867-77

Olson MV, Varki A (2003) Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet 4:20-8

Orti R, Potier MC, Maunoury C, Prieur M, Creau N, Delabar JM (1998) Conservation of pericentromeric duplications of a 200-kb part of the human 21q22.1 region in primates. Cytogenet Cell Genet 83:262-5

Otting N, Doxiadis GG, Versluis L, de Groot NG, Anholts J, Verduin W, Rozemuller E, Claas F, Tilanus MG, Bontrop RE (1998) Characterization and distribution of Mhc-DPB1 alleles in chimpanzee and rhesus macaque populations. Hum Immunol 59:656-64.

Perret M (1997) Change in photoperiodic cycle affects life span in a prosimian primate (Microcebus murinus). J Biol Rhythms 12:136-45

Samonte RV, Eichler EE (2002) Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3:65-72

Schaner P, Richards N, Wadhwa A, Aksentijevich I, Kastner D, Tucker P, Gumucio D (2001) Episodic evolution of pyrin in primates: human mutations recapitulate ancestral amino acid states. Nat Genet 27:318-21

Stankiewicz P, Lupski JR (2002) Genomic architecture, rearrangements and genomic disorders. Trends Genet 18:74-82

t Hart BA, van Meurs M, Brok HP, Massacesi L, Bauer J, Boon L, Bontrop RE, Laman JD (2000) A new primate model for multiple sclerosis in the common marmoset. Immunol Today 21:290-7

Warren KS, Verschoor EJ, Langenhuijzen S, Heriyanto, Swan RA, Vigilant L, Heeney JL (2001) Speciation and intrasubspecific variation of Bornean orangutans, Pongo pygmaeus pygmaeus. Mol Biol Evol 18:472-80.

Yunis JJ, Prakash O (1982) The origin of man: a chromosomal pictorial legacy. Science 215:1525-30.

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A 95:3708-13.

Zhang Y, Ryder OA (2001) Genetic divergence of orangutan subspecies (Pongo pygmaeus). J Mol Evol 52:516-26.

Zimonjic D, Kelley M, Rubin J, Aaronson S, Popescu N (1997) Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. Proc Natl Acad Sci USA 94:11461-65

IV. Detailed Proposal to Identify Additional Human Genetic Variation (Component 3): Sequencing Additional Human Genomes

The following was an integral part of the overall plan by the Working Group to annotate the human genome. The Coordinating Committee and NACHGR were favorably disposed to the concept but wished to consider in more detail the various strategic and technological issues surrounding the proposal. Accordingly, NHGRI has convened a workshop to discuss these and related issues before deciding how to proceed.

A. Background and Rationale

Discovery of human genetic variation and its relationship to human susceptibility to disease is a major priority of the NHGRI and the NIH more broadly. Whole-genome sampling of multiple humans has the potential to extend and complement current approaches to the study of human genetic variation. These approaches largely involve two strategies:

1. PCR-based resequencing of specific genes and classes of genes

Genes are targeted for resequencing either in reference or case-control populations. Data on reference populations define the common alleles and provide some information about the extent of linkage-disequilibrium between them. Typically, they also establish approximate allele frequencies in the major human sub-populations. Variants discovered in reference populations can then be used in genotype-phenotype-association studies of case-control populations. Alternately, the PCR-based resequencing can be carried out directly on case-control populations. The latter approach has the advantage that particular case-control populations may be enriched for variants that are rare in reference populations.

Strengths. Because it can be targeted to specific genes, this method has provided a "fast track" for genotype-phenotype studies of genes that are known or suspected to be involved in particular human diseases. There have been clear successes—examples include identification of variants in several genes that influence progression of AIDS (Silverberg *et al*. 2004) and the discovery that mutations in the *MC4r* gene account for a few percent of individuals with extreme obesity (Lubrano-Berthelier 2003). A major advantage of these methods is that targeted resequencing is well suited to deep surveys of variation in selected populations; hence, it has the potential of correlating rare, as well as common, alleles with phenotype.

Although several initiatives have explored the practicality of extending this approach to all genes, few of the resultant data are in the public domain. One of the more promising approaches to whole-genome sampling involves coupling somatic cell genetics (which provides haploid templates), long-range-PCR amplification, and chip-based sequencing (Patil *et al*. 2001). However, most experience with this method—as with more conventional approaches to whole-genome surveys—has been in the private sector and has produced only proprietary databases.

Weaknesses. With the exception of the somatic-cell-hybrid/chip-based-sequencing approach, targeted resequencing has largely been limited to surveys of exons, exon-intron junctures, and known regulatory regions. The main exceptions are small genes, which can be resequenced in their entirety. Automated processing of raw data acquired from diploid templates remains a problem. Without extensive manual curation, the frequencies of false positives and false negatives are unknown. Insertion-deletion polymorphisms are difficult to score reliably and can mask other variation present in heterozygous individuals. The inherent variability with which different amplicons can be analyzed would be a major problem if PCR-based methods were scaled up to the level required to produce a definitive catalog of human variation.

2. Development of the techniques and infrastructure required for whole-genome-allelic-association scans

In this strategy, variants are used as genetic markers rather than as candidate causal mutations. The basic idea is to genotype case-control populations for a set of SNPs that has a high likelihood of including markers in linkage disequilibrium with the phenotypically important variation. Major initiatives are underway to identify a suitable set of SNPs, to characterize the level of linkage disequilibrium between these SNPs and other nearby variation, to lower the cost of genotyping SNPs, and to establish robust statistical methods for assessing the significance of observed associations (The International HapMap Consortium 2003).

Strengths. This strategy shares the virtue of other "positional cloning" methods of providing a generic approach to establishing genotype-phenotype correlations that does not depend on prior knowledge of underlying biological mechanisms. Since the functionally important variation is to be discovered on a case-by-case basis, implementation requires only a light initial sampling of human variation. This sampling simply serves to identify an adequate number of genetic markers.

Weaknesses. Since present activities are largely developmental, it remains to be seen how well this strategy will work. There are both methodological and biological uncertainties. Methodologically, the number of markers required and the size of study populations both remain poorly defined. Genotyping costs are dropping, but the projected costs of single-phenotype-whole-genome scans remain high. Biologically, the extent of genetic heterogeneity underlying complex human phenotypes remains largely unknown. If many disease phenotypes can be caused by mutations in a wide variety of genes, statistical power will be low. Even when a single gene has a major effect, statistical power will be weak if allelic heterogeneity is high since different causal alleles will be on different haplotypes. Finally, some segments of the human genome will be inaccessible to this approach because patterns of linkage disequilibrium are too complex and short-range to capture with a practical number of genetic markers.

The case for deep, systematic whole-genome sampling of human variation

Introduction. Broadly viewed, the case is the same as that for reference-genome sequencing. It would be far more efficient to define all common human genetic variation in a systematic effort than to rely on piecemeal accumulation of data. All targeted approaches are inefficient, have uncertain and variable quality control, and involve guesswork about what to target. This guesswork has already led to massive acquisition bias in our knowledge of human-genetic variation.

In the public sector, the only substantial dataset that does not suffer from major acquisition bias is that produced by the SNP Consortium. In December 2003, the public trace repository included approximately 10^7 sequencing reads, which provide on the order of 2X coverage of the haploid human genome. There is not yet a detailed published analysis of the full data set. However, at an earlier stage in the sampling—at which point roughly half the reads were available—1.4 million SNPs were identified (International SNP Map Working Group 2001). It has been estimated that this collection contained about a quarter of all SNPs with a minor-allele frequency > 0.4% and a little over 10% of those with a minor-allele frequency > 1% (Kruglyak and Nickerson 2001). Since that time additional sequence has been deposited, including some from the Celera effort, bringing the total coverage to ~7 fold coverage. On the order of 8 million SNPs have been detected from this data.

The SNP Consortium initiative was designed to discover genetic markers, not to characterize human variation. An initiative to achieve the latter goal would need to have a much larger scale. For example, it would require resequencing the equivalent of approximately 100 haploid genomes to have a 95% probability of discovering any SNP with a minor allele frequency > 1% (Kruglyak and Nickerson 2001). Why would these data be useful?

1. They would complement human-primate comparisons in identifying human genes that have experienced unusual patterns of selection during human evolution. Human-primate comparisons allow identification of genes that have experienced strong directional selection. Intra-species variation allows detection of genes that have experienced diversifying selection or recent selective sweeps. The former leads to more, and the latter to less, than typical neutral levels of variation.

2. By definition, they would encompass nearly all common variation in the human genome. To the extent that common variants play a major role in predisposition to common phenotypes, the actual causal variants would be known. Of course, the problem of relating specific genotypes to specific phenotypes would remain; however, many new paths to the discovery of such relationships would be opened. To cite a simple example, it would be possible to develop genotyping assays for all variants that appear—on the basis of our general knowledge of gene function—to be null or strongly hypomorphic mutations. Such assays could then be applied in parallel to diverse case-control populations.

3. In the process of exhaustively sampling common variation, major insights would be obtained into rare variation. For example, it is likely that loss-of-function mutations in most human genes are viable as heterozygotes. Indeed, mouse-knockout data suggest that homozygous loss-of-function mutations are viable in a solid majority of human genes. However, we have little idea what the allele frequencies of such mutations are in most human genes. The incidence of typical recessive genetic diseases suggests that a reasonable guess would be $10^{-3} – 10^{-2}$. In the process of discovering essentially all common variants, a substantial proportion of these rarer alleles, which are particularly likely to be functionally important, would be discovered.
4. A program to discover most common human variants would put genomics on a path toward direct discovery of genotype-phenotype correlations through whole-genome sampling of case and control individuals. Eventually, such methods will almost certainly displace high-throughput genotyping. The path towards implementing them is likely to involve experimentation with a variety of cheap sequencing methods that will typically produce data of poorer quality (i.e. shorter reads and higher error rates) than current technology. Interpretation of these data would be greatly facilitated if we had reliable knowledge from established methods of all common human variation.

B. Full Proposal

The principal method for surveying human variation would be whole-genome sampling of individuals who are representative of the major human sub-populations. Although theoretical efficiencies in SNP discovery could be gained through pooling strategies, these modest efficiencies would not compensate for the attendant loss of information. The best approach would be to sample individuals of known geographic origins. The major issues to resolve would be the depth of the sampling and the number and choice of human sub-populations to be sampled.

Depth of sampling involves an inherent tradeoff between completeness and redundancy. Pooling strategies attempt to minimize redundancy by defining completeness in terms of a particular sub-population rather than its component individuals. However, most of these benefits can be achieved, at insignificant added cost, by light sampling of a series of individuals. The efficiency with which new base pairs are surveyed for variation falls off rapidly as the sampling redundancy on each individual increases. For example, at 0.1X sampling (based on a haploid genome size), the efficiency of an individual-by-individual survey is 93% as high as that of sampling from a deep pool of individuals, whereas at 1.0X sampling it has fallen to <50% of the efficiency of the pool-based strategy. Of course, at such low sampling levels, most of each individual's genotype remains undefined. As whole-genome studies of human variation become established as an approach to analyzing phenotypic variation, this limitation will become unacceptable. However, in the years immediately ahead, it would be reasonable to focus on the aggregate rate of variant discovery rather than to optimize analysis of particular individuals.

In rough numbers, an initiative to discover most common human variation via 0.1X sampling of a series of individuals would require analyzing 1000 individuals. A project of this scale would provide 100X coverage of the haploid human genome and require approximately 320 Gbp of raw-sequence data. A reasonable approach would be to allocate a certain fraction of sequencing capacity to this initiative on an ongoing basis. The fraction could rise over time since the discovery and characterization of human variation is the most open-ended of current proposals to utilize sequencing capacity. If current capacity is ~100 Gbp/yr, allocation of one-third of capacity to the variation initiative would allow half the data to be collected in 5 years, even assuming a flat level of data acquisition. Since the proposal involves random sampling, half the proposed data set would already provide many of the benefits of the full set, whose scale has been chosen somewhat arbitrarily. On the basis of the results of the initial 5-yr effort, follow-up planning would have the benefit of vastly improving our knowledge of human variation.

References

International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933.

Kruglyak, L. & Nickerson, D. A. *Nature Genet* **27**, 234-236 (2001). Variation is the spice of life.

Lubrano-Berthelier C, Cavazos M, Dubern B, Shapiro A, Stunff CL, Zhang S, Picart F, Govaerts C, Froguel P, Bougneres P, Clement K, Vaisse C. Molecular genetics of human obesity-associated MC4R mutations. *Ann N Y Acad Sci*, **994**, 49-57 (2003).

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719-1723 (2001).

Silverberg MJ, Smith MW, Chmiel JS, Detels R, Margolick JB, Rinaldo CR, O'Brien SJ, Munoz A. Fraction of cases of acquired immunodeficiency syndrome prevented by the interactions of identified restriction gene variants. *Am J Epidemiol* **159**, 232-241 (2004).

The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796 (2003).