Center Initiated Proposal:

Expanding Comparative Genomics in Nonhuman Primates

Summary:

We will generate *de novo* whole genome assemblies and associated information about genetic variation (SNPs) for seven nonhuman primate species. These species were chosen through discussions between BCM Human Genome Sequencing Center faculty and the Genome 10K international consortium project. The goal of Genome10K is to facilitate and co-ordinate the sequencing of 10,000 vertebrate genomes over the next several years. This effort is intended to sample phylogenetic and adaptive diversity in vertebrates, including nonhuman primates. HGSC has selected these seven primate species (four Old World monkeys, two New World monkeys and one strepsirrhine) because they add significantly to the phylogenetic coverage across primates, and because they include species with particular biological characteristics that will add valuable new information to our understanding of genome evolution and adaptation. The HGSC is committed to driving continued progress in comparative genomics, with emphasis on primates. This CIP requests approval for expansion of this comparative primate genomics in co-ordination with the Genome 10K project.

Background:

The goal of this Center Initiated Project is to advance the field of comparative primate genomics by sequencing additional primate species designated as significant new targets through discussion and collaboration with the Genome 10K consortium project. The BCM Human Genome Sequencing Center has outstanding prior experience with *de novo* whole genome assemblies of insect and mammalian genomes. We have completed the assembly of the rhesus macaque, and are near completion and publication of three more primate genomes: marmoset, gibbon and baboon. We also have other active mammalian sequencing projects (e.g. deer mouse and dolphin).

There are three fundamental reasons why comparative primate sequencing is valuable to the biological research community. First, primate whole genome DNA sequences are critical for reconstructing and interpreting the processes that produced the modern human genome. Second, genome sequences for nonhuman primate model organisms are valuable information resources for biomedical research concerning human health and disease. Finally, as researchers in basic biology and more disease-oriented research work to understand the mechanisms that generate DNA sequence variation, the impact of natural selection on that sequence variation, and the best strategies for detecting the genomic signatures of historical episodes of natural selection, comparative primate sequencing in species that have undergone well-documented periods of natural selection for particular identifiable traits can be useful analogs for evidence of selection in humans. Continuing

improvement of methods for detecting DNA sequence targets of selection depend on access to additional case studies for the development and evaluation of new analytical tools. The seven species we propose to sequence in this CIP include primates that will contribute to all three of these avenues of research. We propose investigating four Old World monkeys: drill (*Mandrillus leucophaeus*), patas monkey (*Erythropcebus patas*), gelada (*Theropithecus gelada*), and black and white colobus (*Colobus* sp.). We also propose two New World monkeys (white-fronted capuchin, *Cebus albifrons*, and buff-headed capuchin, *Cebus xanthrosternos*) and one strepsirrhine (sifaka, *Propithecus verreauxi*). Except for a few gibbon species, all extant ape (hominoid) species have been sequenced or are now in progress. Consequently, further progress in the study of comparative primate genome structure, content and evolution will come from developing additional breadth of information in primate clades that are more distantly related to humans. As we will explain below, further genome assemblies for Old World monkeys, New World monkeys and strepsirrhines will contribute to our understanding of both the evolution of the human genome and to biomedical genomics (see Figure One for species phylogeny).

The four Old World monkeys proposed here include three cercopithecines (drill, gelada and patas) and one colobine (black and white colobus). The colobus project constitutes an important extension of genome sequencing to a major unstudied branch of primate phylogeny. The primary dichotomy within the Old World monkeys is the divergence of the colobine monkeys from the cercopithecine monkeys (approximately 16-19 million years ago), and to date all the Old World monkeys sequenced, approved or in process are cercopithecines. Thus, production of a whole genome assembly for the black and white colobus will significantly expand coverage of the primate phylogenetic tree by including a major new branch. Furthermore, colobines are uniquely adapted for leaf-eating and hind-gut fermentation, and thus make an interesting biological (physiological and genomic) contrast to all other primates. The three cercopithecines include one species that is important for future progress in HIV-AIDS related research and two that provide windows into rapid and adaptively significant natural selection. The drill (*Mandrillus leucophaeus*) is a natural host for a lentivirus closely related to HIV. Understanding how primates that are natural hosts for lentiviruses are protected from adverse effects of infection is one critical element of current HIV research (Brenchley et al. 2010, Immunity 32:737). Drills also represent a new branch within cercopithecine evolution that have undergone significant morphological evolution.

The two other cercopithecine primates proposed here are important representatives of primate diversity. Both the patas and gelada exhibit major adaptive changes involving ecological and physiological specializations. The patas monkey is closely related to African green monkeys, but has undergone substantial changes in anatomy and behavior, and therefore provides an opportunity to study the signatures of evolutionary adaptation and anatomical change at the genomic level. The same is true for the gelada, which is related to *Papio* baboons, but has evolved remarkable and unique specializations related to diet, dentition, locomotor and postural behavior, communication behavior and other traits. Compared to the closely related but less specialized *Papio* baboons, geladas have undergone rapid evolutionary change in many biological systems. Assemblies of the genomes of African green monkeys and *Papio* baboons are complete, and will be published

soon, so access to whole genome assemblies for geladas and patas monkeys will open a wide range of opportunities for comparative genomics and the analysis of the origin of new evolutionary specializations.

The sequencing of whole genomes for capuchins (genus *Cebus*) is justified for several reasons, and consequently the Genome10K consortium has targeted these two species of capuchins for priority sequencing. First, New World monkeys are under-represented among primates currently approved or in progress. Only the marmoset and squirrel monkey are currently under study, despite the wide diversity present among New World monkeys. Second, the genus *Cebus* is composed of an extraordinary series of species. Production of whole genome assemblies for these two species will help identify anthropoid or primate specific regulatory elements, as a result of their substantial divergence from humans, apes and Old World monkeys. However, other benefits will also accrue from the sequencing of these two capuchin monkey genomes. Unlike essentially all other non-ape species, capuchins are skilled users of a variety of tools. Their unusual ability to fashion useful tools from natural objects, and to manipulate their environment to gain access to preferred and otherwise unobtainable foods has drawn significant attention from primatologists. The evolutionary development of such cognitive skills, complex learning and locomotor and manual manipulative abilities in a New World monkey was a surprise, and therefore further investigation of capuchins at a variety of levels is well justified.

One crucially important goal of comparative primate genomics is to facilitate broad comparisons among distantly related genomes, in order to increase power to detect evolutionary conservation of DNA sequences outside conserved protein coding genes. Construction of a whole genome assembly for strepsirrhine primates such as Verreaux's sifaka will significantly increase our ability to detect sequence conservation among primate taxa. The consensus estimate for the date of evolutionary divergence separating strepsirrhines (e.g. the sifaka) from humans is 60-65 million years, with some estimates even older. Currently, there is no draft genome assembly published for any strepsirrhine, though one of these species (the bushbaby, genus *Otolemur*) has been sequenced and awaits publication. Our addition of the sifaka (which is not particularly closely related to the bushbaby) to the panel of available primate whole genome sequences will provide valuable analytical power to studies of sequence conservation and the reconstruction of the ancestral primate genome.

Preliminary Data/ Samples:

We have identified specific sources for all the seven DNA samples required to produce reference genome sequences. Oliver Ryder (San Diego Zoo, and one of the organizers of Genome10K) has tissue and/or DNA samples from the white-fronted capuchin, gelada, drill, patas, and black and white colobus immediately available in his Frozen Zoo. Dr. Anne Yoder (Director of the Duke Univ. Primate Center) has agreed to provide materials for Verreaux's sifaka. Dr. Hector Seuanez (Federal Univ. of Rio de Janeiro) will provide samples for the buff-headed capuchin.

Experimental Plan:

Sequence production for these genomes will use methods that are now standard and well established at the BCM-HGSC. The primary approach will be to use the BCM-HGSC Illumina Hi-Seq pipeline, but we will also evaluate the utility of newer technologies as appropriate. We will generate 80x Illumina paired-end small insert sequence coverage for each reference genome, and a further 80x coverage in Illumina mate-pair reads (1kb, 3kb, 5kb and 8kb inserts). We will also explore the use of the Pacific Biosciences RS instrument to generate low coverage of long (> 2-4 kb) continuous reads that may be effective in closing gaps within scaffolds.

For genome assemblies, we will employ available Next Gen assembly methods and produce a preliminary assembly for each genome. Iterative scaffolding and gap filling steps using the Atlas system modules Atlas-Link and Atlas-GapFill will be used to improve the initial assemblies prior to release. As part of the Assemblathon consortium, we are actively involved in evaluating various Next Gen assembly methods, having used a variety of assemblers to produce initial or intermediate files in the Atlas assembly system. The current best contender for initial WGS assembly is ALLPATHS-LG, with SOAPdenovo also a potentially useful tool. ALLPATHS-LG and SOAPdenovo have different expectations and requirements for input libraries (e.g. non-overlapping vs. overlapping paired-ends reads), and these issues influence our choice of methods and strategy. We have experience with both systems, as well as with Phrap, Newbler, CABOG and others. This experience indicates that the optimal choice for assembly engine depends on the details of the data available and the characteristics of the genome. In light of frequent new releases of updated versions of software, we plan to remain flexible concerning assembly strategy, but will initially employ the Atlas system in combination with ALLPATHS-LG, and continue evaluating other options. Where appropriate tissues are available, we will perform RNA sequencing for up to four tissues using the Illumina pipeline. State-of-the-art tools for analyzing RNA sequence data (Tophat, Cufflinks, Bowtie, other new developments) will be used to characterize the transcriptome of the species for which RNA sequence data can be produced.

As an integral part of the Experimental Plan, we will identify and obtain DNA samples from additional individuals from each of the seven species to pursue discovery of intra-species genetic variation. The number of additional individuals sequenced will vary across these seven species due to differences in sample availability. Our goal is to produce 30x Illumina read coverage for each of 4-8 animals per species. We will use these data to generate SNP lists, identify CNVs and structural variation, and investigate other aspects of genome variation (e.g. polymorphic Alu insertions). Reaching a sample size of eight individuals will be relatively easy for species that are more numerous in captivity, such as colobus and sifaka. It will be more challenging for others (e.g. patas and buff-headed capuchin), but the HGSC faculty have strong relationships with the primate research community (both the people who manage colonies of captive primates and field researchers) and we expect to be able to obtain satisfactorily numbers of samples to generate valuable data on within-species variation. The Frozen Zoo at the San Diego Zoo, the Genome10K consortium and

the American Society of Primatologists are all valuable resources we will use to fill out panels of samples for each species.

We anticipate that the cost of this Experimental Plan will be approximately $146,000 per species. This is based on an estimate of $106,000 for production of deep sequence data for a reference animal, lower sequence coverage for eight additional individuals, and RNA-seq for four tissues. We also include $40,000 for required labor for bioinformatic analysis and genome assembly.

Data Release/Timeline:

We plan to have samples for five species (white-fronted capuchin, drill, gelada, black and white colobus and patas) in our hands before the end of December 2011, and we will begin library construction at that time. For the sifaka and buff-headed capuchin, we will work with Drs. Yoder and Sueanez to obtain samples during the first quarter of 2012, with laboratory activities beginning soon after arrival of appropriate materials. Sequencing of these seven genomes will proceed as quickly as possible, with 2-3 undergoing sequencing simultaneously. We expect to complete all the sequencing of these seven species during the second quarter of 2012, and to produce a series of initial assemblies before the end of 2012.

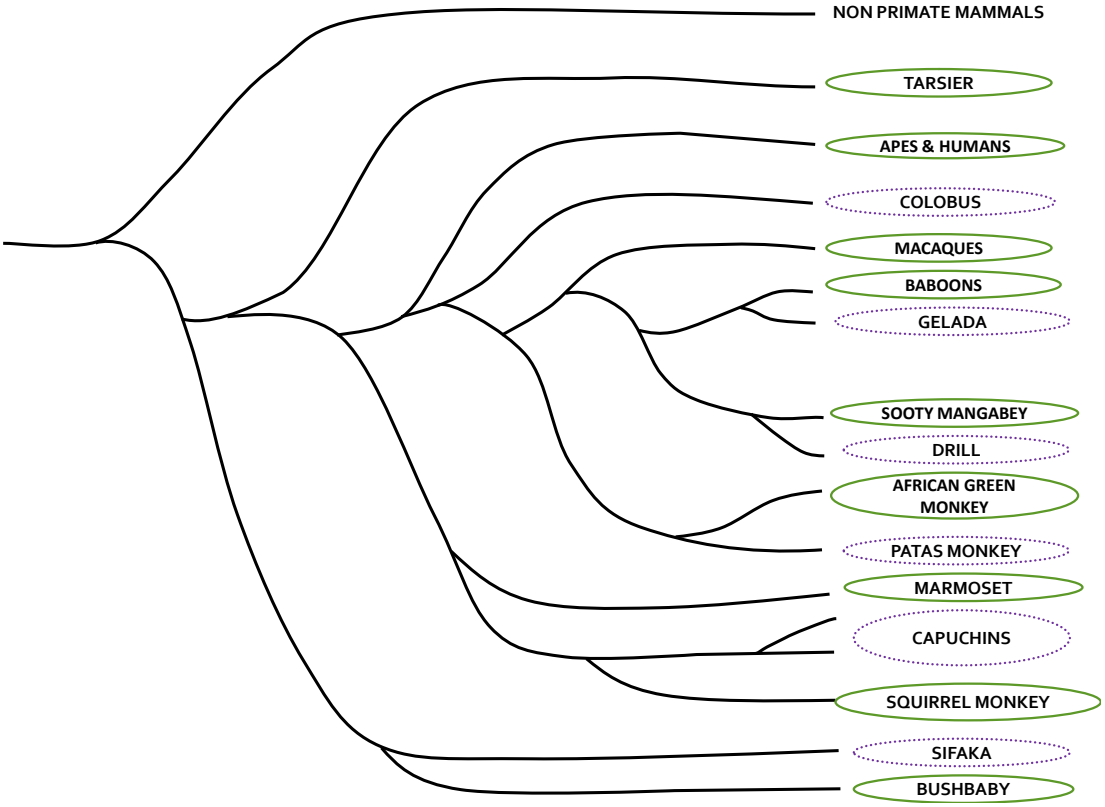

Figure One: Phylogeny of Selected Primate Species. The seven species proposed for sequencing in this CIP are shown in dotted ovals, while the primate genomes that have been or are now being sequenced are shown in solid green ovals.