Re: Proposal for BAC library construction of Orangutan (*Pongo pygmaeus)*
Date: November 15, 2001
From: Evan Eichler, Ph.D., Dept. of Genetics, Case Western Reserve University
To: BAC Library Resource Network, National Human Genome Research Institute

Importance: The orangutan is considered as the outgroup hominoid species for most great ape/human genetic studies. 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 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). Recently published SNP studies emphasize the value of genomic sequence from this organism to determine the ancestral and derived status of human alleles (Chen and Li 2001; Kaessmann et al. 2001). This species is particularly valuable when data from gorilla and chimpanzee are ambiguous with respect to the ancestral status of a common human polymorphism. Genetic data from both subspecies of orangutan 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.

The orangutan karyotype is the best representative of the ancestral hominoid ancestral state--both humans and African ape chromosomes are believed to be largely derivatives requiring a minimum of 10-15 chromosomal rearrangements from this hominoid archetype (Muller and Wienberg 2001; Yunis and Prakash 1982). Specific regions of the hominoid genome evolve much more rapidly than "generic" DNA and therefore require a closer primate outgroup species (other than macaque and baboon) in order to resolve the complexity of these regions. Processes such as Y chromosome evolution, pericentromeric duplication, subtelomeric rearrangements and centromere repositioning necessitate the construction of this library. As an example, the orangutan genome is most frequently used to determine the timing and movement of recent segmental duplications associated with chromosomal rearrangement disorders (Velocardiofacial/DiGeorge, Prader-Willi Syndrome, Smith Magenis, etc), pericentromeric duplications and subtelomeric rearrangements. These regions comprise an estimated 5-7% of the human genome and exhibit accelerated rates of evolutionary turnover (Bailey et al. 2001; Consortium 2001; Eichler 2001). 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). The construction of a BAC library is necessary in order to survey the structure and organization of these regions over large expanses of genomic sequence (many of the duplications or sites of rearrangement are in excess of 100 kb). Comparative sequencing will provide insight into the underlying mechanisms that have predisposed to duplication-mediated rearrangements associated with human genetic disease (Lupski 1998).

Usage: The primary use of this library would be for comparative sequencing purposes of targeted genomic regions. It is anticipated that select regions of high biological/biomedical interest (immunological genes, genes under positive Darwinian selection, regions of rapid genomic rearrangement, haplotype characterization, etc) would be primary targets. Due to the relative high degree of genomic sequence identity between orangutan and human (95-96%) it is unlikely that the orangutan BAC library would be used for a complete genomic sequencing effort. The availability of a high quality human reference sequence, however, in combination with a high quality, random orangutan BAC library would foster genome-wide comparisons (BAC-end sequencing or fingerprint overlay against human) in order to identify regions of hypervariability or BAC-end discordancy. It should be noted that a whole-genome shotgun approach (whose inserts are relatively limited in size <10 kb) would be ineffective in the resolution of complex genomic regions as outlined above. Furthermore, despite the low level of sequence divergence (~4%), sufficient variation exists within "generic" DNA to thwart the development of contiguous orangutan sequence over large genomic DNA using PCR methodology . For example, an attempt to analyze eight regions from the orangutan X chromosome (each 100 kb in length) by PCR showed 15% failure after two rounds of oligonucleotide design and amplification based on human reference sequence (Eichler and Chakravarti, unpublished). In addition, to its primary use for comparative sequencing, an orangutan BAC library would be a valuable resource to the cytogenetics research community. BAC clones would be used as probes to investigate the chromosomal evolution of the hominoid species. Such a resource may be ultimately used in conservation studies (CGH microarrays or through cytogenetic comparisons) to more effectively manage the breeding program of this endangered species both among wild and captive populations (Zhang and Ryder 2001).

Research Community: Interest in the development of an orangutan library is broad (immunology, human genetics, genomics, chromosomal evolution, species conservation etc) but currently focused to a specific subset of researchers in each field. Among the human genetics community, this includes researchers interested in recapitulating the evolutionary history of human haplotypes (eg. Aravinda Chakravarti, Svante Paabo, Maynard Olson and Wen-Hsiung Li), those studying recently duplicated regions associated with human genomic disorders and rearrangements (eg. Jim Lupski, Evan Eichler, Barbara Trask and Tamim Shaikh) and those interested in large-scale primate comparative genomic sequencing efforts (eg. Eric Green, Ken Dewar, Bruce Roe and Shaying Zhao). Within the field of immunology, there is a long-standing interest in the development of these resources for the characterization of regions associated with blood group antigens or "hypervariable" genes of the immune system (eg. Peter Parham, Ron Bontrop and Eric Long). Those interested in chromosomal evolution (centromere

repositioning, Y chromosome evolution and comparative primate cytogenetics) (eg. Mariano Rocchi, David Page, Johannes Wienberg and Malcolm Ferguson-Smith) would benefit from such a resource. Finally, a strong interest has been expressed by those in the area of species conservation (Oliver Ryder and Lori Perkins) that such a resource and the concomitant sequence data generated from it would enhance their ability to manage and monitor genetic diversity within wild and captive populations of this endangered species.

Has the organism been proposed to NHGRI or another publicly funded agency for BAC-based genomic sequence? No formal proposal, to my knowledge, has been made. Informal discussions regarding great-ape sequencing projects have been discussed and considered (McConkey and Varki 2000).

Complementary Genomic Resources. A collection of DNA and cell line material from six unrelated members of this species are available from Coriell Cell Repository. Dr. Peter Parham has an expanded resource of lymphoblastoid cell lines from both unrelated and related orangutan cohorts that he has made available upon request. Both cDNA and genomic material are now available from the Center for Reproduction of Endangered Species (Oliver Ryder).

Proposed Strain Selection. Two subspecies of orangutan are recognized: Bornean (*P.p.pygmaeus)* and Sumatran (*P.p.abelli)*. Analysis of mitochondrial and microsatellite DNA suggest that the Bornean subspecies shows the greatest amount of genetic diversity—although the difference is slight (Warren et al. 2001). If there is a choice, it is recommended that Bornean orangutan DNA be obtained.

Genome Size: Genome size is comparable to human, 3 Gb.

Source DNA: Two sources of high quality DNA have been identified. Cell line material from an orangutan male has been obtained from Dr. Peter Parham and has been assessed for gross rearrangement (deletions or duplications) by CGH BAC microarray analysis using a Human panel of 2500 BAC DNAs (in collaboration with Dan Pinkel). A request for a second source of DNA generated directly from blood DNA of a Bornean orangutan male is being pursued through the orangutan Survival Species Plan Coordinator, Lori Perkins.

Specifications: A random BAC library (*EcoR*1 partial digest, insert size (not less than 150 kb), 10-fold genomic redundancy) from a male Bornean orangutan should be considered using standard cloning vector. Libraries significantly less than 150 kb insert will complicate the analysis of duplicated regions as these regions are often in excess of 100 kb. Regions of rearrangement when compared to a human reference sequence require multiple clones to eliminate the possibility of artifacts due to BAC rearrangement. Consequently sufficient depth (10X) is requested.

Time frame: Many analyses of complex regions of the human genome *already* require the existence of this library. Some investigators have been forced to screen inferior cosmid libraries from more distant species (i.e. Gibbon, *Hylobates klossi* library) to begin

to understand the evolution and implications of these regions (Johnson, 2001). The need for this library is immediate and should be among the first libraries sequenced as part of the NHGRI effort.

Other Support: No other support has been requested or is available for the construction of this library.

Other Relevant Information: None.

References.

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

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005-17.

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

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.

The International Human Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-920

Eichler EE (2001) Segmental duplications: what's missing, misassigned, and misassembled- and should we care? Genome Res 11:653-6.

Eichler EE, Lu F, Shen Y, Antonacci R, Jurecic V, Doggett NA, Moyzis RK, et al (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

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

Jackson MS, Rocchi M, Thompson G, Hearn T, Crosier M, Guy J, Kirk D, et al (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 the great apes. Nature in press

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.

Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14:417-22.

McConkey EH, Varki A (2000) A primate genome project deserves high priority. Science 289:1295-6.

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

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, et al (1998) Characterization and distribution of Mhc-DPB1 alleles in chimpanzee and rhesus macaque populations. Hum Immunol 59:656-64.

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 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