Non-human Primate Genome Sequencing
January 2006

General Considerations

The long-term goal of non-human primate genome sequencing has been to annotate the genetic differences that have emerged over the last 60-70 million years of primate evolution. Embedded in these differences are the changes that have shaped evolutionary divergence and produced our unique human traits. Because differences rather than conserved sequences represent the major findings, high quality sequence is required. With the high quality data it will be possible to detect subtle changes in the primary structure of genes, which is important since even single amino acid differences can be key contributors to morphological differences. Similarly the important changes in the sequence of gene regulatory regions may involve only small numbers of DNA bases. Additionally, the important differences can involve large-scale genome rearrangements, including duplications and other structural rearrangements. Scaffolds built from BAC-end sequences and/or BAC skims can facilitate identification of the rearrangements, but cannot solve the pressing question of what are the precise events in divergence of this class of sequence that occurs between closely related species. We therefore advocate that one path to the general improvement and addition of value to primate genome sequences is to specifically select regions known to be involved in duplications and other structural rearrangement for further study. In practice this involves selection of individual large insert clones that capture these events, followed by their additional analysis to near-finished DNA sequencing standards. .
This approach that aims to refine selected parts of primate genomes is different but complementary to the goals stated elsewhere of globally improving the knowledge of rearrangements across whole genomes, and other proposals to refine regions in draft genomes that contain genes. **Overall, the current proposal shares with previous recommendations the general aim of pursuing fewer index primate species sequenced at higher quality, rather than many at low quality.**

Three criteria have been considered in the selection of non-human primate genome sequencing targets: 1) phylogenetic position with respect to human; 2) utility for identification of primate-specific regulatory elements; and 3) biomedical relevance in cases where other factors were equal.

1) Primate Phylogenetic Nodes.

Past work has documented that there are seven distinct phylogenetic branchpoints between human and non-human primate species. Each of these nodes provides considerable power for annotation of the human genome, revealing those elements that have been altered in the evolution of the human species. Multiple species comparisons are expected to be significantly more informative than individual pairwise comparisons that have been possible to date [1]. Complete (~6 X) genome sequencing projects have been approved by Council for four of these seven nodes, represented by chimpanzee, orangutan, macaque and marmoset (Table 1, see Figure 1 below). A fifth branchpoint (Gorilla gorilla) was officially approved October 2005 by the Wellcome Trust (Sanger

Institute). In addition, low coverage genome sequencing projects have been approved for the remaining anthropoid branch, represented by the gibbon (0.1 X BAC end sequence coverage approved by Council 5/24/05) and for bushbaby (galago), a member of the prosimian branch (2X coverage as part of the low coverage mammalian sequencing proposal).

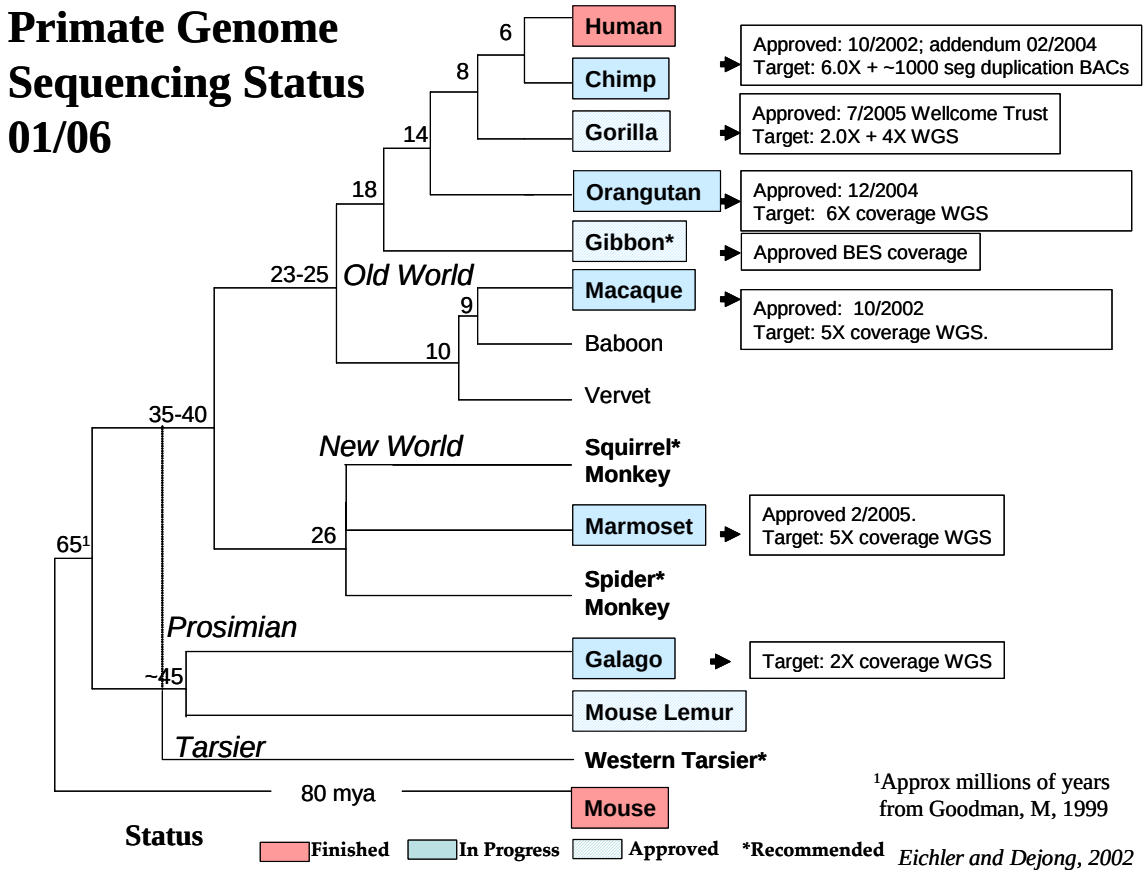


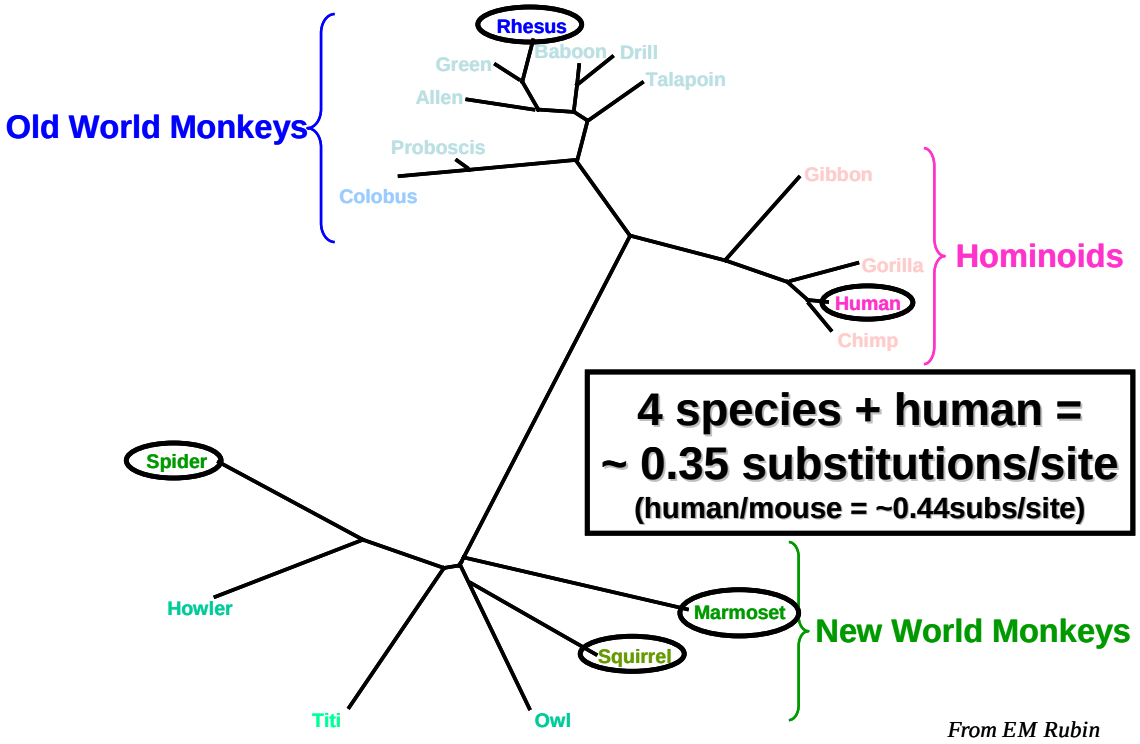
Figure 1. Present and proposed primate sequencing projects

The position of tarsiers on the primate tree is still debated. In a recent analysis of 15 nuclear genes, tarsiers placed as a separate branch among Prosimians with strongest statistical significance [2]. However, other analysis suggests that tarsiers may represent a monophyletic group with New World and Old World Monkey (Haplorrhine hypothesis) and therefore an eighth node with respect to human (Figure 1). Regardless of its position tarsier shows an accelerated substitution rate, similar to galago and mouse lemur (Figure 1, strespsirhines), when compared to other sampled non-human primate genomes due possibly to small body size, high metabolic rates and shorter generation time.

To complete the sampling of the primate nodes, we propose a 6-7X whole genome assembly of the gibbon genome and low coverage sequencing of the tarsier and mouse lemur (the latter is also included in the mammalian sequencing initiative). Together the tarsier, bushbaby and mouse lemur genomes (6X combined coverage) will serve as an outgroup for anthropoid sequence comparisons. The 2X coverage of the tarsier genome will also permit its unambiguous placement on the primate tree.

2) Identifying primate-specific elements

A special group discussion of our working group (including Dr. Svante Paabo and Eddy Rubin) focused on the utility of additional primate genome sequence for the purpose of identifying primate-specific regulatory elements. Both the number of species and the depth of coverage required for *phylogenetic shadowing* were discussed. Although the frequency of such primate-specific regulatory elements is unknown, Dr. Rubin's group [3] has convincingly identified several anthropoid and cattarhine-specific elements based on multiple sequence comparisons. There was a general consensus that additional New World monkey sequences would be most informative, as Old World Monkey sequences offered limited branch length while Prosiman sequences were too diverged and comparable to signal obtained from other non-primate mammal sequencing projects. Phylogenetic shadowing requires a balance between genetic distance (branch length) and proximity for optimal alignments. Relatively few additional species are required. Five species, in addition to human, chosen to maximize the phylogenetic diversity of the sampled primate species, capture 80% of the sequence variation present in the full species dataset. Two additional NWM genomes are proposed (squirrel monkey and spider monkey) at complete sequence coverage. Higher sequence coverage is required for discovery in order to unambiguously place sequence differences in the context of the primate genome (D. Haussler, personal communication). These data in conjunction with full genome sequence data from macaque and marmoset would be used as a basis for evaluating higher sequencing coverage and additional species at a later date possibly as new sequence technology emerges(e.g. technology from 454 Life Sciences) [4,5].



From EM Rubin

Figure 2: Recommended species for phylogenetic shadowing: Squirrel and Spider Monkey at 6X coverage.

3) Primates of Biomedical Relevance.

Baboon (Papio sp.) is one of the few non-human primate models that are well studied for diseases of blood and circulation. A substantial colony of baboons (>4,000) is available at the SFBR Southwest Foundation and have been studied physiologically and behaviorally (John Rogers). In the longer term (6X) WGS plasmid sequencing or the equivalent will be warranted to support biomedical research.

Specific Recommendations

1. Complementary BAC-based sequencing of Complex Regions.

To date most whitepaper proposals and 'deep draft' genome projects have primarily focused upon extensive WGS coverage of genomes. In some cases BAC and fosmid clone sequencing, either via clone end pairs or by light coverage of individual or pooled large clones, has added to the overall data to be accumulated for genome assemblies. In the case of WGS assemblies without BACs there is generally only very poor recovery of duplicated or complex genomic regions. In assemblies with added BAC or fosmid information there has been recovery of some of these regions, but the representation still falls short of that known to be expected in complete genome sequences.

As a result, there is a paucity of data about duplications and other biologically complex regions drawn directly from genome assemblies. Instead, most of our knowledge has come from subsequent or independent 'follow up studies' that examine subsets of the duplicated regions. For example, a recent duplication analysis using whole genome shotgun sequences between chimpanzee and human[6] has shown a) that significant genetic differences lie in such regions (2.7% segmental duplication vs. 1.2% single-basepair differences), b) that many of these regions correspond to rapidly evolving gene families and c) that these regions are not adequately assembled within 3 or 6X drafts[6] (Eichler, unpublished). These regions are among those being targeted as part of the chimpanzee genome "refinement".

As a first priority, we recommend to characterize these duplications and rearrangements more precisely in the course of producing these genome assemblies because they will help us more quickly understand:

(a) Similarity and divergence between primates, providing a more balanced view of human genome variation, including regions of structural variation, segmental duplication and lineage-specific rearrangements,
(b) Orthologous gene relationships between humans and more distantly related mammals allowing patterns/rates of gene deletion/duplication, selection and pseudogenization to be evaluated.
(c) Comparatively rare primate-specific regulatory elements that we suspect to be more abundant in recent duplications.
(d) The duplication of gene loci that are relevant to infectious disease, drug response and/or vaccine development in both human and model systems, e.g., genes important for immunity (i.e. alpha, beta defensins, chemokine ligand receptors, HLA, etc) and drug detoxification (cytochrome P-450 gene families, glutathione S-transferases, carboxylesterases, etc), which are particularly enriched within these complex regions of the genomes [6].

In order to add to the value offered by WGS studies that are already underway we therefore recommend more refined sequencing of BACs selected from available assembly data, representing these regions of potential duplication or other structural rearrangement. **We specifically recommend the detailed analysis of up to 1,000 BACs per index project (macaque, marmoset and orangutan).** This should then become a standard for future draft primate genome projects.

2. New Projects and Additional Sequence. Additional WGS sequence of four primate species are also recommended for the purpose of either better representation of phylogenetic nodes for human genome annotation, for generating preliminary data for the purpose of phylogenetic shadowing, or for their role in biomedical studies.

A. Gibbon (Nomascus Leucogenys). Recommendation: 6X WGS sequencing (5.5 X plasmid and 0.5 X fosmid). This is the last remaining catarrhine node for which no sequence has yet been generated. The lesser apes represent the link between human/great apes and the Old World monkey species. Its sequence provides a unique view of evolutionary divergence from the human genome over 18-20 mya of species separation[7,8]. BAC-end sequencing has been approved, representing 0.1 X sequence coverage. BAC-end sequencing has proceeded slowly due to delays in the distribution of the gibbon BAC library. Costs for fosmid end-sequences have dropped considerably over the last 6 months (D. Smith personal communication). Structural variation of the gibbon genome is thought to be extensive. Paired-end sequence from 0.5 X fosmid sequence (15 fold physical coverage) in conjunction with 0.1 BAC end sequence (10-fold) would provide a high quality overview of the gibbon genome prior to sequencing. Paired-end sequencing mapping [9] would simultaneously identify and subclone most regions of large-scale rearrangements. Depending on the complexity of the region either fosmid or BAC clones could be subsequently selected for higher quality sequence analysis to resolve the structures of the rearrangements. Both cell line material and blood may be obtained from the same female gibbon used for BAC library construction (Dr. Alan Mootnick at the Santa Barbara Zoo, CA, Director of Gibbon Conservation)

B. Tarsier (*Tarsier bancanus*) Western Tarsier. Recommendation: 2X WGS plasmid sequencing**.** Combined with bushbaby and mouse lemur the addition of this species will provide sufficient sequence coverage (6 X) to reconstruct an outgroup for anthropoid sequence comparisons. A recent analysis suggests that prosimians consist of three extant groups (Lorisiformes, Lemuriformes, and Tarsiiformes). The genetic branch length for any pair (~0.20-0.25 substitutions per site) would be sufficient to allow reconstruction of an outgroup prosimian genome by combining these three sets of data. WGS data from tarsier would also clarify whether this species represents an additional phylogenetic node with respect to human. If so, deeper coverage sequencing may be warranted for this species. Frozen material is available from the Duke Primate Center, but this is likely unsuitable for BAC and fosmid library preparation. Additional sources of material are being sought. .

C. Squirrel Monkey (*Saimiri bolilviensis*): Recommendation: Complete (6X) WGS plasmid sequencing for the purpose of phylogenetic shadowing (discussed above). Phylogenetic shadowing requires the identification of high quality basepair differences (as opposed to conserved elements) within an unambiguous orthologous genomic context. Whole-genome assembly of 6X sequence coverage allows most sequence errors to be effectively eliminated and provides sequence contigs of sufficient length for this purpose. The squirrel monkey is a member of the New World Monkeys, estimated to have diverged from the anthropoid common ancestor (35-40 mya). Genetically, at least six anciently separated New World monkey clades are recognized that diverged from each other ~ 20 mya [10]. Squirrel monkey represents a different branch than marmoset. It is the second most commonly used NWM in biomedical research (4451 Pubmed references for genus Saimiri). Bolivian squirrel monkeys (*Saimiri boliviensis)* are most frequently used for malaria vaccine studies. A BAC library (CHORI-254) exists for *Saimiri boliviensis* and material is available from Dr. Larry Williams, University of South Alabama, National Squirrel Monkey Breeding and Research Resource (SMBRR).

D. Spider monkey (*Ateles geoffroy*i): Recommendation: Complete (6X) WGS plasmid sequencing for the purpose of phylogenetic shadowing (discussed above). It is selected as another distant branch of the NWM. Biomedical: One of the few small mammal models for hepatitis B virus research. A small colony of black-handed spider monkeys (*Ateles geoffroyi*) is available at the SFBR Southwest Foundation (Robert Lanford).

3. Overall Upgrading of non-human Primate Genomes

The long term ambition of the program should be to provide genome sequences for all these non –human primates that are of the standard we have reached for human. The group recommends that the possibilities for upgrading primate genomes, beyond the additional BAC characterizations suggested above, are systematically and continually evaluated. The template for this evaluation should be the accompanying document and guidelines on ‘genome refinement’ When the costs and efforts for the refinement of additional mammalian genomes is acceptably low then the same list of primate sequences presented here should be further sequenced and refined.

Concluding Remarks

The benefits of primate genome sequencing will be greatly enhanced once high-quality draft sequence becomes available. Simply overlaying sequence reads against the human genome reference, while an interesting exercise, humanizes our genomic view of our most closely related species. The focus of non-human primate genome sequencing is variation as opposed to conservation. This emphasis requires that assembly artifacts be eliminated. It is tempting to propose to draft many more NHP genome sequences of lower quality as opposed to generating higher quality sequence assemblies of existing NHP genome projects. Experiences with the chimpanzee genome assembly indicate that this would not be wise. In this light, the highest priority in 2006 should be the improvement of the chimpanzee, macaque, orangutan and marmoset by the addition of the BAC sequence data described above. Next the additional primates can be sequenced to similarly high standards. In the longer term the complete genomes of these key non human primates can be refined and other primate sequences can follow.

References

1. Chimpanzee Sequencing and Analysis Consortium.A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* (2005).
2. Eizirik, E., Murphy, W.J., Springer, M.S. & O'Brien, S.J. Molecular Phylogeny and Dating of Early Primate Divergences. in *Anthropoid Origins* (eds. Kay, R. & Ross, C.) (New Visions, 2004).
3. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391-4 (2003).
4. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).
5. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-32 (2005).
6. Cheng, Z. et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88-93 (2005).
7. Goodman, M. The genomic record of Humankind's evolutionary roots. *Am J Hum Genet* **64**, 31-9 (1999).
8. Muller, S., Hollatz, M. & Wienberg, J. Chromosomal phylogeny and evolution of gibbons (Hylobatidae). *Hum Genet* **113**, 493-501 (2003).
9. Newman, T.L. et al. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15**, 1344-56 (2005).
10. Chiu, C.H. et al. Reduction of two functional gamma-globin genes to one: an evolutionary trend in New World monkeys (infraorder Platyrrhini). *Proc Natl Acad Sci U S A* **93**, 6510-5 (1996).