Increasing sequence coverage from 2x to high coverage (6-7x) for selected mammalian species
January 2006

Summary

Background

The nomination and implementation of low coverage (2x) whole genome sequencing of 16 mammal species by NHGRI in 2004 and 2005 has provided a broad comparative database for genome analysis and human genome annotation.   Low coverage sequence has indeed proved efficient for recognizing features of the human genome shared across most mammals. However, for other analyses, there are significant limitations of low coverage genome analyses, due primarily to the fact that it provides only about 80% of the genome sequence, leaving several hundred thousand gaps. The limitations include:

First, the gaps create interrupted and missing genes in the non-human species, making it impossible to define complete genes in non-human species for the purpose of comparing them to the human genes. Such comparative analysis is a major way in which we learn about human genes. This is especially important, for example, in the investigation of paralogous gene families, which often have different numeration, location, organization, and presumed function in different mammal lineages. Functionally related gene clusters whose proximity, order, and organization are adaptive (e.g. the major histocompatibility complex, cytokine, cytokine receptors, chemokines, chemokine receptors, immunoglobulins, Killer immunoglobulin genes, Toll-like and T-cell receptor gene clusters) are disappointingly abridged and disorganized in low coverage sequences.

Second, the occurrence of segmental duplications (estimated at 5% of the human and 1-2% of the mouse genome) is nearly impossible to annotate with low coverage. These regions are important in the generation of new genes by duplication. Deeper WGS sequence allows such regions to be readily identified (Bailey, 2002; Cheng, 2005) and improves assembly within duplicated regions, particularly where genomic sequence identity falls below <97% (Eichler, mouse genome analysis, unpublished).   However, we should note that even at 6-7X coverage large (> 15 kilobases) and highly identical (> 97%) duplications may not be adequately resolved (She, 2004). These features require large insert clone data for resolution.

Third, it is impossible to precisely define the large-scale chromosomal rearrangements (Murphy et al., 2005) that have shaped the history of the human genome using only 2X coverage genomes. Genome assembly of the developed 2X sequence, with interrupted contigs in the hundreds of thousands is highly dependent upon, and thereby intrinsically biased by, the comparative synteny of genes assembled within the high coverage or "finished" species' assemblies producing a distorted view of the actual syntenic situation. Further, inversions and reciprocal translocations are frequent enough to play an important role in human genetic disease, and have also left an important mark on the evolution of the genome.

Fourth, copy number polymorphisms (CNPs), endogenous retroviral elements, or common insertion/deletion (indel) variants cannot be annotated adequately with low coverage.

All of these genomic features, descended to human and other mammal species from a common ancestor, are critically important elements in the human genome and their interpretation would richly benefit from more highly accurate genome sequences.

Sequencing Objectives for high coverage genomes

We see three principal objectives for which higher quality sequence data from a limited set of mammals would significantly aid in the annotation and understanding of the human genome. These are: 1) the identification of the “core” mammalian genome; 2) the increased utilization of medically relevant mammals; and 3) the reconstruction of the ancestral eutherian genome. We describe the rationale for each in more detail below and propose specific species. Based on the experience with the mouse, dog and cow genomes, 6-7X fold total shotgun coverage should provide the required assembly quality, with N50 supercontig sizes in the tens of megabases, more than 97% of the assembled bases with an error rate of less than 10^{-4} and some 95-98% of the genome in supercontigs. Two-fold coverage is already available or has been proposed for each of these species in the context of the low coverage mammalian sequencing proposal, significantly reducing the sequencing demands.

1. The ‘core’ mammalian genome, consisting of those regions that are common to all mammals. A major goal of the human genome annotation is to recognize short conserved sequence blocks that are evolutionarily conserved due to functional constraints, likely structural, genic, and network regulatory elements (Margulis et al., 2005; Eddy, 2005). Within mammals, these comprise the core mammalian genome sequences, i.e. those conserved across multiple ordinal lineages of mammals and constituting a complete complement of the "core" genome sequences utilized across mammal differentiation.

The deletion of large segments of unconstrained regions of the genome during the evolution of mammals provides an important resource for identifying the ‘core’ mammalian genome. Intersecting the human, mouse and dog genome already restricts attention to about ~812 Mb. By obtaining deep coverage of a few additional mammals, it should be possible to continue to narrow this core genome. A key question is whether the core genome is substantially larger than the 5% of the human genome (roughly 150 Mb) identified by nucleotides under purifying selection.

Such analysis requires deep coverage (rather than 2x coverage), because inferences must be drawn from the absence of sequence in a genome. The most efficient approach is to focus on species that have had the most extensive deletions of the ancestral eutherian genome. Evidence suggests that the deletion rate is roughly linearly correlated with branch length. (Dog has a branch length of 0.22 and has deleted ~21% of the ancestral genome, while mouse has a branch length of 0.41 and have deleted ~44% of the ancestral genome. Although large deletions cannot be assessed for the 2x mammals, smaller deletions do show proportionality.)

The importance of branch length is seen by the following comparison:

- With 4 dog-like genomes, the core genome should be reduced to ~410 Mb.
- With 4 mouse-like genomes, the core genome should be reduced to ~210 Mb.

(This analysis assumes a core genome of 150 Mb, but the basic point holds under any assumption.)

In selecting species for this objective, we specifically sought ones that had a relatively fast genome-wide evolutionary divergence, thus maximizing the accumulated sequence divergence from the human baseline (sum = 0.956 substitutions/site from human; see figure). Four species are recommended:

1. Lesser hedgehog tenrec	*Echinops telfari*
2. Nine-banded armadillo	*Dasypus novemcinctus*
3. Guinea pig	*Cavia porcellus*
4. European common shrew	*Sorex araneus*

2. Utilization of animals with 'biomedical' value and potential model species. These species have particular advantages with multiple homologs for human hereditary, infectious and chronic diseases plus the potential for informing human medicine due to technological advances as model species. Each of these species also represents distinctive mammal lineages, but do not add as much divergence as those proposed for objective 1 (sum = 0.501 substitutions/site from human; see figure). Four species are recommended for high sequence coverage based upon biomedical criteria:

1. Domestic cat	*Felis catus*
2. Brown bat	*Myotis lucifugus*
3. Horse	*Equus caballas*
4. Rabbit	*Oryctolagus cuniculus*

White papers for these species have been submitted, and a brief summary of the specific advantages of each are provided in an Appendix to this report.

3. The 'reconstruction' of the ancestral eutherian genome, including the specific key evolutionary changes that led from it to the human genome. Four super-ordinal mammal clades (Afrotheria, Xenarthra, Euarchontoglires, and Laurasiatheria) have been recently resolved, providing a foundation for the systematic study of the human genome from its roots in the common ancestor of all placental mammals (Murphy et al 2001a;b). A relatively brief period of extraordinary fecundity and evolutionary success created the wonderful diversity of mammals that exists today. Reconstructing the key evolutionary changes that occurred in the human lineage during this period distinguishing our heritage from that of the other mammals, demands that we obtain an accurate picture of the ancestral structure of the genome before each major available divergence node. Non-human species are used as outgroups in this analysis, to allow reconstruction of the ancestral state. For this purpose, it is most useful to sequence slowly evolving species rather than rapidly evolving species needed for objective 1. In Figure ?, we identify 12 major divergence nodes that reach back to the common ancestry of all placental mammals. We identify six nodes (and nominated species) within primates, two for Euarchonta (primates, tree shrews and flying lemurs), and four ancestral nodes that capture Euarchontoglires, Boreoeutheria (Lauasiatheria plus Euarchontoglires), Xenarthra and Afrotheria.

The proposed primate genomes are considered in the separate Primates Report from the AHGWG. The indicated primate species (Figure ?) have all either already been sequenced or have been proposed for high coverage sequencing for other reasons or by other groups. Six non-primate eutherian species are recommended for high sequence coverage based upon reconstruction criteria:

1. African savannah elephant	*Loxodonta africana*
2. Tree shrew	*Tupaia spp.*
3. Nine-banded armadillo	*Dasypus novemcinctus **
4. Brown bat	*Myotis lucifugus **
5. Rabbit	*Oryctolagus cuniculus **
6. Horse	*Equus caballas**

* also in group 1 or 2.

In addition to the large scale changes that occurred during the evolution of the human genome, over this evolutionary period it is possible with sufficient outgroup data to reconstruct specific ancestral states down to the level of individual bases with an estimated 98% accuracy (Blanchette et al, 2004). Even though many of them are low coverage, the existing and proposed genomes in should suffice to achieve this base-level accuracy in ancestral reconstruction. Such data would be of extraordinary value in recapitulating the dynamic history of the genomic steps that occurred during human evolution, as well as across the mammalian radiations in general. *In the context of annotating the human genome, the precise evolutionary history of a functional genomic element greatly aids in its initial identification within the genome sequence, yields vital clues to its function, and provides an important context in which to study that element in terms of human variation and disease.*

Species recommended by AHGWG .

In sum the above objectives argue for a total of ten species for high (6-7x) coverage starting in Spring 2006. Reasons for inclusion are given in relation to the three criteria defined above: ‘core’, ‘biomedical’, and ‘reconstruction’:

1. Lesser hedgehog tenrec	*Echinops telfari*
2. Nine-banded armadillo	*Dasypus novemcinctus*

These two represent unique super-ordinal mammal clades which are unrepresented by a highly accurate sequence. Tenrec has the most rapid genome divergence in its clade, Afrotheria, an advantage for understanding the ‘core’ genome. (Figure 1-Core).

3. Domestic cat	*Felis catus*

The cat is a major biomedical model, with an active large veterinary community that has described over 200 hereditary homologues, infectious disease models for SARS, AIDS, leukemia/sarcoma, bird flu, and chronic disease (See Appendix for details)

4. Brown bat *Myotis lucifugus*

Bats represent 20% of all mammals, a distinct lineage with developing genetics models for gene transfer based genetic manipulation. (See Appendix for details)

5. Guinea pig *Cavia porcellus*

Value in both ‘core’ as rather rapid divergence as a new rodent lineage and also modest biomedical value (Figure 1-Core).

6. African savannah elephant *Loxodonta africana*

Value in both ‘reconstruction’ and ‘biomedical’ model, particularly neuroscience (Figure 2; Appendix)

7. Horse *Equus caballas*

Primarily Biomedical, but represents an unrepresented placental mammal order Perrisodactyla (Figure 2; Appendix)

8. European common shrew *Sorex araneus* (Figure 1-Core)
9. Tree shrew *Tupaia spp* (Figure 2- Reconstruction)
10. Rabbit Oryctolagus cuniculus (Biomedical-Appendix)

Additional considerations for five species selected (cat, elephant, bat, horse and armadillo). In order to maximize the human genome annotation as well as to increase utility of biomedical species for genetic and other research, the AHGWG recommends that a portion of the genome sequence centers appropriation be dedicated to three important genomic resources, not included in the original 2X sequence proposal. The three supplements include:

1. **Development of cDNA sequence data to complement gene annotation and interpretation in these model species.** There is near universal recognition that the most effective data to enable the direct annotation of genomic DNA sequence is accurate full length cDNA sequence since the two can be easily aligned to represent the functional gene. Likewise, computational prediction of gene sequences can be supported and verified by overlapping real cDNAs. The complete cDNA sequences greatly exceed the value of EST data because of the continuity that is provided. This is particularly useful

when the genome assembly is fragmented, as is the case with draft genomes, since genes are often interrupted by gaps.

The cDNAs should be from libraries constructed in order to maximize the likelihood of recovering full-length clones. Several tissues should be sampled per species and 5 and 3 EST sequencing of clones from each library should be followed by complete coverage of the additional bases in selected cDNAs at a high accuracy. More discussion may be required for each of these parameters but 5 tissues per species, with 10,000 ESTs from each and a subsequent total 10,000 cDNAs analyzed with a final cumulative error of not more than 1/5,000 bases would represent appropriate aims. The cDNA clones themselves should be made available for distribution but it is not necessary to develop highly curated clone collections as the primary value will be in the sequence data.

2. **Y-chromosome targeted sequencing from BACS and fosmid library developed from flow sorted Y-chromosomes in five species.** The human Y-chromosome has been sequenced to high quality, but so far no other mammalian Y-chromosome has received high level of sequencing (Skaletsky et al., 2003). Preliminary comparative sequencing data exists for the chimp, which shows some interesting differences form the human Y-chromosome in terms of functional gene conservation (Hughes et al., 2005; Rosen et al 2003). A portion of the mouse and chimp Y has been sequenced along with a number of Y-chromosome genes from mouse, chicken other organisms. Yet Y chromosome in nearly all mammal species is largely an un-chartered genomic area, even Drosophila Y-chromosome sequence is unknown. The human Y-chromosome includes about 28 Mbp - approximately 1% of the genome - including at least 80 protein coding genes, eight massive palindromes and a smorgasbord of gene conversion recombination in the MSY (male specific Y-chromosome) region.

   The Y-chromosome specifies the most obvious human (or mammalian phenotype (our gender) and triggers all things sexual. Knowledge of Y-chromosome genomics has informed forensics, human anthropology, chromosome evolution, species phylogeography, sex-hormone influenced cancers, gameto-genesis and reproduction. Nonetheless the Y-chromosomes small size, haploid number in males, and other considerations have led to the selection of females for the 16 species nominated for whole genome sequence assessment (2X coverage). The AHGWG believes that the Y-chromosome should not be neglected further, rather be included as a target of draft sequencing for five species here recommended for elevation to 6-7x sequence coverage **(cat, elephant, bat, horse and armadillo).**

   One plausible approach would be similar to the strategy employed for human, chimp and mouse, using a combination of Y chromosome specific BAC sequences and fosmid clone sequences derived from flow sorted Y-chromosome derived fosmid libraries. BAC libraries and metaphase chromosome sorts of most of these five species have already been achieved emphasizing the feasibility of the strategy.

3. **SNP development across the species by sequencing 12-20 individuals selected to capture genetic diversity within the species**. Genetic markers, e.g., STRPs, SNPs and deletion/insertion polymorphisms (DIPs), are valuable resources for understanding the

biology of a given species. The human genome has been subject to numerous polymorphism discovery efforts, e.g., The SNP Consortium and the International HapMap Consortium, along with large-scale genotyping, e.g. Perlegen and HapMap (International HapMap Consortium, 2005). These efforts have enabled a multitude of researchers to better utilize the human genome for disease association mapping and population genetics. Similar benefits can be realized for other species given the availability of a polymorphism resource.

In the case of nominated out-bred species, (elephant, bat, horse and armadillo), a significant number of polymorphisms will be discovered from the 6-7 fold coverage of the single individual. For example, judging from the SNP discovery rate of about 1 SNP per 2kb of sequence analyzed from the 2X elephant sequence, the total number of SNPs discovered within this single elephant individual at 6-7 fold coverage should total 1.5M SNPs. However, since this is a single individual, the representation of SNPs will be biased away from rare SNPs and towards more common SNPs. Of common SNPs above the 5% frequency, these 1.5M SNPs will only represent about 20% of all SNPs within the population of these elephants. To generate additional markers in the homozygous regions we would need to sequence additional unrelated individuals. These polymorphism levels seen in the elephant genome are expected to be several-fold lower than expected in bat, horse and armadillo: the elephant, *Loxodonta africana*, has a remarkable reduction in overall genomic variation relative to other mammals (Roca et al, 2001). In contrast to these species, the domestic cat is inbred to a significant extent and as such will require specific WGS from unrelated cats from other breeds to achieve a useful level of SNP coverage, as was required for the dog SNP discovery (Lindblad-Toh et al 2005).

References

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S,
Adams MD,Myers EW, Li PW, Eichler EE.
Recent segmental duplications in the human genome.
Science. 2002 Aug 9;297(5583):1003-7.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa
K, Church D,DeJong P, Wilson RK, Paabo S, Rocchi M, Eichler EE.
A genome-wide comparison of recent chimpanzee and human segmental
duplications.Nature. 2005 Sep 1;437(7055):88-93.

Eddy, S. R. A model of the statistical power of comparative genome sequence analysis.
PLoS Biol. 2005 Jan;3(1):e10. Epub 2005 Jan 4.

Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature. 2005 Sep 1;437(7055):100-3.

International HapMap Consortium, 2005 A haplotypes Map of the Human genome nature 437:1299-1320.

Lindblad-Toh et al.Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005 Dec 8;438(7069):803-19.

Margulies EH, Vinson JP, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, Green ED, Lander ES, Mullikin JC, Clamp M; NISC Comparative Sequencing Program. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A. 2005 Mar 29;102(13):4795-800. Epub 2005 Mar 18.

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.

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 radiation using Bayesian phylogenetics. Science 294: 2348-2351, 2001.

Murphy, W.J., Larkin, D.M., Everts-van de Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., Hitte, G., Meyers, S.N., Ostrander, A.E., Pape, G., Parker, H.G., Raudsepp, T., Rogatcheva, M.B., Schook. L.B., Skow, L.C., Welge, M., Womack, J.E., O'Brien, S.J., Pevzner, P.A., Lewin, H.A.: Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309:613-617, 2005.

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

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. Abundant gene conversion between arms of palindromes in human and ape Y-chromosomes. Nature. 2003 Jun 19; 423(6942):873-6.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003 Jun 19;423(6942):825-37.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM,
Sutton G,Halpern AL, Eichler E. Shotgun sequence assembly and recent segmental duplications within the human genome.
Nature. 2004 Oct 21;431(7011):927-30.

Appendix Justification for Biomedical Model species

A. Domestic cat

White Paper
http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/CatSEQ.pdf

1. Cat is a major medical model for over 200 human hereditary diseases. The 2x sequence has already led to the mining and implication of hereditary model for neurological and retinal atrophy disease and disease gene identification will benefit substantially from increased sequence coverage.
2. There is an established cat genome annotation group already assigned 18 aspects of the feline genome which relate to human annotation .
3. Of the species selected for 6-7x and 2x coverage, the cat has the most conserved (un-rearranged) genome relative to all the other mammals (see Murphy, et al., 2005). This means that cat and human genome organization represent the only index ancestral genome arrangement, critically important for human and mammal genome annotation and evolutionary inference.
4. Cats have medical model of many human deadly infectious diseases notable, FIV-AIDS; FeLV-leukemia; FIPV-SARS, bird flu, panleukopenia and neurological diseases.
5. Clinical assessment of dogs and cats happen in parallel in world's veterinary community. The third International Conference on Advances in Feline and Canine Genomics convenes in August, 2006
6. Segmental duplication and SNP assessment with the 2x sequence is very difficult and incomplete.
7. Reproductive research, tools and implementation in cats are much advanced with artificial insemination, IVF, embryo transfer and cloning ("CC") already achieved (Shin et al., 2002). Development of embryonic stem cells suitable for therapy research plus KO and transgenic cats is advancing very rapidly.
8. Cats represent one of the most exquisitely successful predators ever evolved, a striking adaptation mimicked and even more successfully achieved by human evolutionary predominance.
9. Over 40 breeds of domestic cats and 37 Felidae species provide a rich opportunity for genomic research in domestication (i.e. history, tracks of LD, recessive allele expression and other aspects) and into species adaptation in a mammalian species group (Johnson et al., 2006).
10. There are an estimated 70 million cats in the USA and many more in other countries, providing plenty of opportunity for sampling and experimental development. Also the short generation time (1-2 years) means that developing informative pedigrees of virtually any genetic character, disease or genetic manipulation can be done in a few years.
11. The feline genetic community has developed appreciable genomic resources including somatic cell hybrid and RH panel and maps over 30,000 tissue specimens, linkage pedigree similar to CEPH, full annotated mtDNA and MHC genome sequence, flow sorted chromosomes for ZOO-FISH, BACs, PAC, fosmid, and cosmid libraries. Y-

chromosome fosmid libraries from flow sorted cat Y chromosome will also be developed shortly.

B. Microbat- Myotis lucifugus - Little Brown Bat –

From a biological perspective, bats are high priority from several perspectives. First the bat family Chiroptera comprise nearly one quarter of all mammals (977 species). Bat species represent important human health related natural reservoirs for rabies, SARS, and probably Ebola. Vampire bat saliva has anti-stroke therapeutic value and they are the only mammals that fly and echolocate. Further, their genome is compact estimated at 2 GB, off-loading at least one third of the genomic material found in human and other species (For this reason alone, identifying dispensable genomic residue is imperative for reconstruction perspective).

As with all bats, **- *Myotis lucifugus*** has an extremely high metabolic rate, yet contrary to the traditional inverse correlation between longevity and metabolic rate and, *M. lucifgus* has been recorded as living for at least 34 years in the wild. This is the longest documented life span of any bat species.

- This bat has a relative long branch length useful for 'core' annotating the human genome.
- It is a representative of the Yangochiropteran subordinal group.
- Myotis lucifugus is a vespertilionid bat (most speciose bat family) and is found in the United States.
- It is an echolocating insectivorous bat, important to the bat community as a whole.
- Vespertilionid bats have on average an extremely small genome size. (average genome size = 2.3 pg, # species =21) Mario Capecchi has estimated the genome size of Myotis lucifugus as 2.0 GB.
- This bat is capable of extended hibernation.
- The Myotis genus has the widest distribution of any bat genus.

There are approximately 4,800 extinct mammalian species of which almost a quarter are bat species. Only rodents outnumber bat species and two rodent species have already been fully sequenced. The diversification of bat species is quite remarkable as are the variation of habitat to which bats have adapted. Economically, the global importance of bats as insect predators and plant pollinators is undisputed.

Among bat species *Myotis lucifugus*, the little brown bat, is the most extensively studied. This species is found all over North America, so availability is no problem. Synchronous embryos and primary fibroblast cell lines are developed. The chromosome number is 22 and comparative cytogenetic reciprocal chromosome paints have established *Myotis lucifugus* as carrying a very highly conserved ancestral organization similar to only human and cat.

Finally, a provocative new genetic system developed by Mario Capechhi will allow researchers to functionally assess any mammalian genome at an *in vivo* physiological level. The strategy allows the transfer of chromosome segments across the entire bat genome into a live mouse thereby allowing phenotypic functional assessment of varying portions of the bat genome (or any other genome similarly manipulated). A collection of mouse lines each containing a different small chromosomal fragment of a chosen mammalian species would contain representation of the entire genome of bats and demonstrate phenotypic variation relative to control mice.

C. Horse *Equus caballas*

1. Biomedical: Several aspects of the physiology and pathology of the horse are highly relevant to human biology and medicine: Severe Combined Immunodeficiency Disease, a fatal genetic disease of Arabian horses), virology (equine infectious anemia retrovirus), zoonoses (viral encephalitides such as West Nile Virus and others), neurology (Cerebellar Abiotrophy and others), muscle disorders (glycogen storage diseases, periodic paralysis and others), musculoskeletal diseases, connective tissue/skin disorders (Junctional Epidermolysis Bullosa and others), reproduction, orthopedics, and respiratory biology. A full genome sequence would support these and other equine research studies with direct application to human health.
2. Equine diseases: The equine genome sequence would also be applied in studies of equine disease and physiology for the benefit of the horse. Gene discovery and quantitative trait loci analyses are paramount in this endeavor. This would include investigations of simple inherited autosomal and sex linked conditions and complex diseases that have both genetic and environmental components.
3. Extensive breeding records: Horse breeders have maintained meticulous pedigree records for many horse breeds, some extending back over 300 years. Deep pedigree records coupled with excellent health and performance records provide excellent material for (QTL) studies. The cooperation of horse breeders with participants in the horse genome workshop has been excellent.
4. The horse is proposed as the representative species of the order Perissodactyla. No other species from this order have yet been sequenced.
5. Community: An organized international community of basic research scientists and veterinary clinicians work in the area of horse genetics. Scientists from over 25 laboratories meet regularly and have been collaborating on equine genome definition and application for the past decade.
6. Tools developed during the last 10 years include synteny, linkage, cytogenetic, RH, comparative and integrated maps, 3 BAC libraries, collections of EST data and several databases and websites for sharing information (http://www.uky.edu/AG/horsemap/).
7. Economics. The applications of the horse genome sequence would benefit the equine industry, which has a yearly economic impact of $102 billion Gross Domestic Product (GDP) and 1.4 million Full Time Employees (FTE) in the United States alone. The population size of domestic horses is estimated at 9.2 million in the United States and 115 million worldwide.

D. European Rabbit *Orychtolagus cuniculus* -

1. Impact on human disease studies: The European rabbit is a major medical model used in toxicology, embryology, physiology, metabolism, immunology and host-pathogen interactions. The rabbit has long been established as a validated animal model for scores of human disease conditions. A complete rabbit genome would extend current rabbit models of human disease. Understanding these disease conditions on a molecular level requires complete sequence-based resources for identifying not only the genes involved in each model, but also pursuing aspects of gene regulation and mechanisms of transcriptional control.
    a. Cardiac disorders: lipid metabolism, atherosclerosis, and cardiomyopathy
    b. Tumor biology: malignant conversion of rabbit papillomas and lymphocytic leukemia in transgenic c-myc rabbits.
    c. Endocrine diseases: acromegaly, diabetes and obesity
    d. Autoimmune diseases: reactive airways disease (asthma), arthritis, and systemic lupus erythematosus (SLE)
    e. Infectious diseases: M. tuberculosis (only the rabbit host models all stages of human tuberculosis), anthrax, coccidioidomycosis, cryptococcosis, candidiasis, enterohemorrhagic E. coli, bacterial meningitis, endocarditis, syphilis (rabbit is the only useful experimental animal model of human syphilis), HTLV-1, and herpes simplex (HSV).
    f. With a complete rabbit genome, gene expression experiments, gene therapy and transgenics could be evaluated in all of these models to define the genetic basis of disease or understand and manipulate host responses.
2. Advantages of the rabbit animal model that will be significantly advanced by a deep coverage rabbit genome:
    a. Transgenics: gene deletion (knock-outs) and transgene insertion (knock-ins) transgenic mutants are now possible through electrofusion of cumulus cells into an enucleated oocyte to generate cloned rabbits
    b. Immunology: lagomorph leukocyte markers, markers associated with cell trafficking, chemotactic molecules, cytokines, T-cell receptors, immunoglobulins, MHC antigens, blood groups, complement and genetic deficiencies, autoimmune diseases and tumors of the immune system are under active research. Given the similarity of disease to that of humans, the ability to interrogate the host-pathogen interface could have important impact on reducing morbidity and mortality in humans with vaccine and drug discovery.
    c. Drug development and testing: comprehensive genetic information for the rabbit will allow cross-species comparison with mouse and rat models to better evaluate of the effect of individual drugs especially in critical areas such as embryofetal development
3. Practical advantages of the rabbit animal model over animal models with complete genome sequence:
    a. Rabbits produce high quality, high affinity antibodies from serum.

b. Rabbits are large enough to permit non-lethal monitoring of physiological changes, prolific, widely available, and easy to handle.

c. The rabbit is an attractive animal model for cardiovascular research. Rabbits have a slower heart rate and larger heart (compared to mice) that allows physiological analyses with echocardiography and cardiac catheterization. Ischemia-reperfusion studies in heart, spinal cord and brain also utilize rabbits

d. The rabbit retina is the premier model for research on the mammalian retina because it is avascular, relatively easy to maintain in culture. Drug safety and toxicity for the eye are also performed with rabbits because of anatomically larger eyes and the response of the ocular surface to drugs that is similar to that of humans

4. Orychtolagus cuniculus is the most well studied species within the order Lagomorpha. No other species from this order have yet been sequenced.

5. Rabbit resources:

   a. A largely closed colony of rabbit lines have been developed, bred, and characterized at NIAID, NIH. These lines maintain polymorphisms of a variety of genes involved in immunity, including genetic variants (allotypes) of the VH, CH, and CL regions of antibody molecules. A relational database of more than 40 years of breeding records is maintained for the colony. The colony also contains descendants of rabbits formerly at the Basel Institute for Immunology that maintain mutations in key immunological loci.

   b. There are 45 recognized breeds of domesticated rabbits that have been bred for different sizes, coat colors and density, ear sizes and shapes, and hair lengths, providing a resource for genetic studies to understand the genetic basis of these heritable traits. Moreover, these breeds may provide the basis for QTL mapping of heritable disease phenotypes as rabbits are prolific and have a short generation time (<1 year) thus characteristics of interest can rapidly be bred and studied.

   c. Inbred rabbit colonies exist with researchers in the Netherlands (Utrecht), and in Japan. Outbred rabbits bred for homozygosity at loci of interest and transgenics have been successfully engineered.

   d. Two BAC libraries are available: LBNL (~7.7x; average insert size 175kb) Therapeutic Human Polyclonals (~4x; average insert size of 124 kb).

5. Rabbit community:

   a. Basic research - 383 grants in the CRISP database utilize rabbits in their research for non-anitsera-producing or biological testing purposes. Over 50 individual investigators wrote letters of support for the rabbit white paper, outlining how their research would be directly affected by a deep coverage rabbit genome.

   b. Husbandry rabbits are prolific and can inexpensively make protein of high quality; 20% of consumed food can be turned into edible meat. Mapping quantitative trait loci (QTL) for fertility, viability and growth are of clear global interest since rabbits are efficient converters of plant fodder (including cellulose rich plants) to edible protein. An international community is interested in the rabbit genome for its contribution to rabbit as a food source (World Rabbit Science Association http://www.dcam.upv.es/wrsa/english/index.htm)

Appendix II; **African savannah elephant.**

1. The elephant is an ideal species for **reconstruction of historic genomic events** (Roca and O'Brien, 2005; Springer et al., 2003); it represents Afrotheria, the most basal (107 Mya) of the eutherian clades, yet at the same time has a shorter branch length than the "core" afrothere, the tenrec. The elephant also represents the clade within Afrotheria most divergent from that of the tenrec (80 Mya), and is therefore the ideal "reconstruction" complement to the "core" tenrec genomic sequence for annotation of the human genome and for characterizing the most ancient superordinal events in the evolution of eutherian genomes.
2. Sizable research community and **biomedical importance** (Roca and O'Brien, 2005; Sukumar, 2003). Elephants are among the most intensively researched wild taxa. They comprise the best-studied order within Afrotheria; even before the genomics era there were more scientific papers and DNA sequences generated for elephants than for all other afrotheres combined. Elephants have a unique morphology with many derived structures, in contrast to the primitive morphology of the other selected afrothere, the tenrec. Their long life spans (70 yr) may be of interest for longevity research, and they are subject to diseases with homologues in humans or livestock, including anthrax, herpesvirus, orthopoxvirus and tuberculosis. Elephants have been used to study endogenous retroviruses and SINEs. Elephant cytogenetic chromosome paints, an RH map, and a BAC library are currently available or in development. Recent technical breakthroughs in ancient DNA that involved extinct relatives of elephants relied on the elephant genome sequence as a comparative standard.
3. **Evolution of sex chromosomes** and cytonuclear interactions (Roca and O'Brien, 2005). In some savannah elephant populations a "conplastic" cytonuclear genomic pattern is present, in which nuclear and mitochondrial genomes reflect distinctive evolutionary histories, suggesting that the elephant may prove useful for studies of cyto-nuclear interactions. Reproductive success in male elephants depends largely on body size, and the cytonuclear pattern results from low reproductive success for small hybrid (savannah x forest elephant) males in competition with larger savannah bulls. Large old savannah males can also out-compete smaller young savannah males, leading to a generation time that is much longer in male elephants than among the less competitive females. Elephants thus represent an opportunity for studying how sex differences in reproductive patterns affect genomic patterns, especially among X- and Y-linked genes. The elephant is a candidate for **Y-chromosome sequencing**, both to examine these effects and because the elephant serves as representative of the Afrotheria.
4. The elephant genome may shed light on the **evolution of advanced traits** (Roca and O'Brien, 2005; Roth and Dicke, 2005); elephants have large brains and cortical volumes that allow them to exhibit relatively high levels of learning and memory, and complex systems of communication and social interaction (e.g., elephants are the only non-human animal in which four hierarchical tiers of social organization have been rigorously

demonstrated). Since these traits in humans likely result from combination and enhancement of properties found in non-human animals, rather than from unique properties, comparative studies may be especially appropriate. Elephants comprise one of only three mammalian lineages (along with primates and cetaceans) in which these advanced traits are present and could be studied on a comparative basis. They are therefore also candidates for **cDNA sequence datasets** to be generated for brain tissues, although testicular tissue may also be of interest since elephant testicles remain undescended (testicondy) in adult males.

5. The savannah elephant genome sequence will facilitate development of additional nuclear markers for **conservation genetics** useful across the three extant (and endangered) species (Roca and O'Brien, 2005). Genetic markers have already identified forest and savannah African elephants as separate species, identified distinctive populations of African and Asian elephants as conservation priorities, and determined that unusual cytonuclear genomic patterns exist among elephants. The elephant would be a leading candidate for **SNP development by sequencing multiple individuals** selected to capture genetic diversity. SNPs should prove useful in conservation, for identifying genetically distinctive populations and for establishing the provenance of poached ivory.

References:

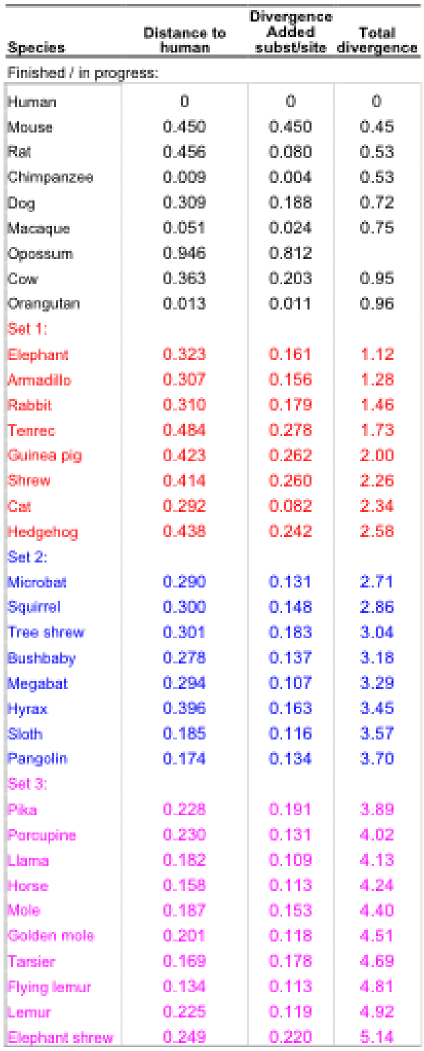
Roca, A. L. and S. J. O'Brien: (2005) Genomic inferences from Afrotheria and the evolution of elephants. Curr Opin Genet Dev 15:652-659.

Roth G, Dicke U: Evolution of the brain and intelligence. (2005) Trends Cogn Sci 9:250-257.

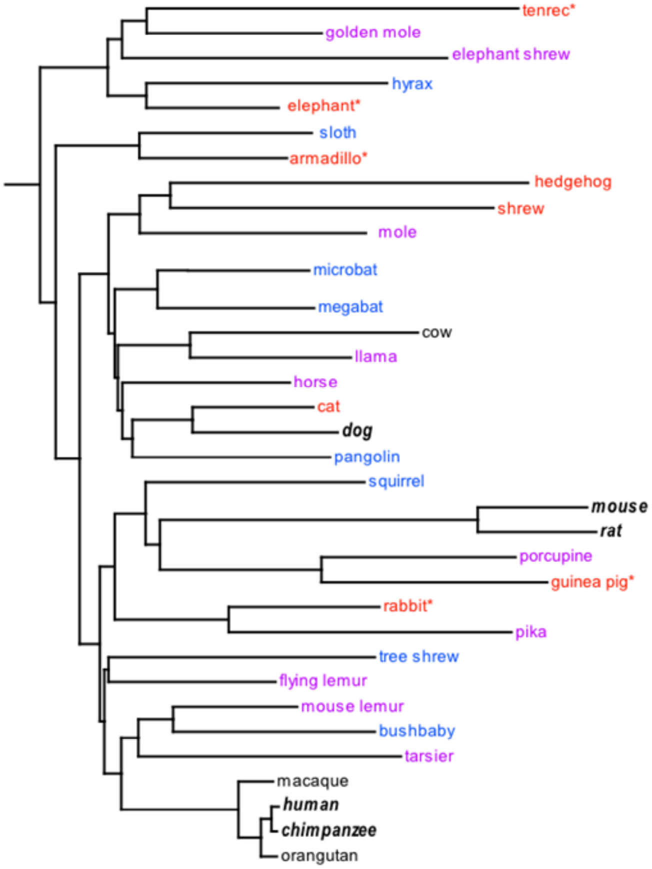
Shin T, Kraemer D, Pryor J, Liu L, Rugila J, Howe L, Buck S, Murphy K, Lyons L, Westhusin M.: (2002) A cat cloned by nuclear transplantation. Nature 415:859.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ: (2003) Placental mammal diversification and the Cretaceous–Tertiary boundary. Proc Natl Acad Sci USA 100:1056-1061.

Sukumar R: (2003) The Living Elephants: Evolutionary Ecology, Behavior, and Conservation. Oxford: Oxford University Press.

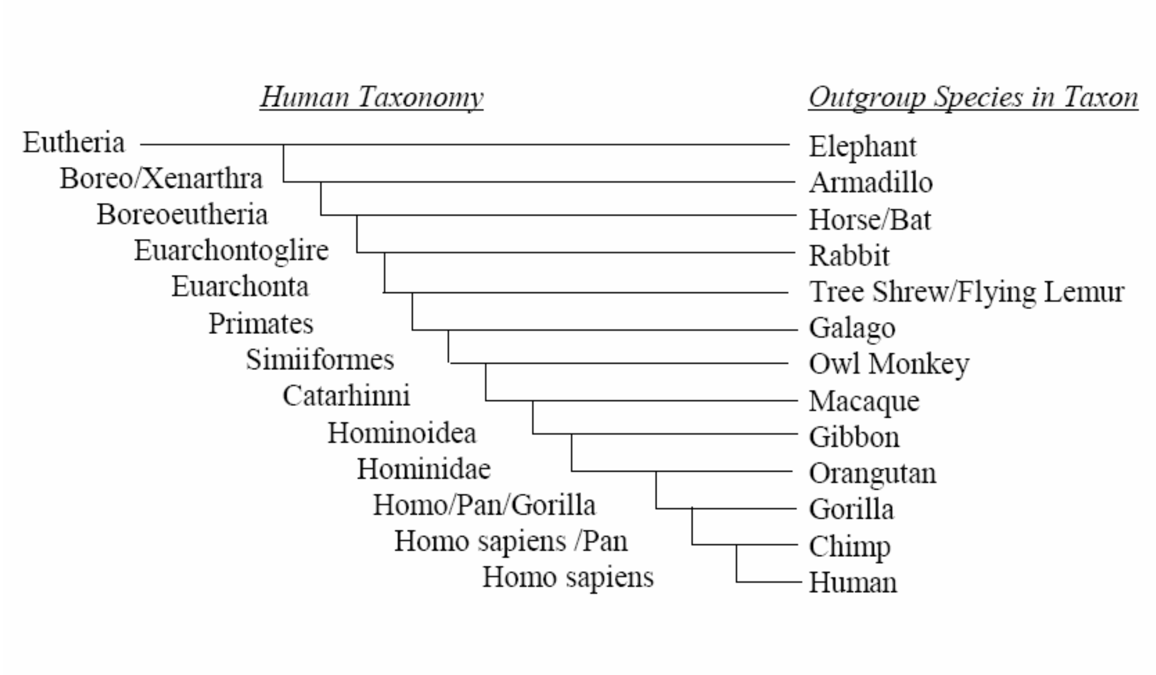


Species	Distance to human	Divergence Added subst/site	Total divergence
Finished / in progress:			
Human	0	0	0
Mouse	0.450	0.450	0.45
Rat	0.456	0.080	0.53
Chimpanzee	0.009	0.004	0.53
Dog	0.309	0.188	0.72
Macaque	0.051	0.024	0.75
Opossum	0.946	0.812	
Cow	0.363	0.203	0.95
Orangutan	0.013	0.011	0.96
Set 1:			
Elephant	0.323	0.161	1.12
Armadillo	0.307	0.156	1.28
Rabbit	0.310	0.179	1.46
Tenrec	0.484	0.278	1.73
Guinea pig	0.423	0.262	2.00
Shrew	0.414	0.260	2.26
Cat	0.292	0.082	2.34
Hedgehog	0.438	0.242	2.58
Set 2:			
Microbat	0.290	0.131	2.71
Squirrel	0.300	0.148	2.86
Tree shrew	0.301	0.183	3.04
Bushbaby	0.278	0.137	3.18
Megabat	0.294	0.107	3.29
Hyrax	0.396	0.163	3.45
Sloth	0.185	0.116	3.57
Pangolin	0.174	0.134	3.70
Set 3:			
Pika	0.228	0.191	3.89
Porcupine	0.230	0.131	4.02
Llama	0.182	0.109	4.13
Horse	0.158	0.113	4.24
Mole	0.187	0.153	4.40
Golden mole	0.201	0.118	4.51
Tarsier	0.169	0.178	4.69
Flying lemur	0.134	0.113	4.81
Lemur	0.225	0.119	4.92
Elephant shrew	0.249	0.220	5.14



CORE Tree FIGURE 1

Maximum likelihood phylogenetic tree of species recommended for whole genome sequence with limb lengths indicated in the Table. (after Murphy et al 2001,a;b; Margulies et al 2005)



Reconstruction Tree FIGURE 2

Anthropocentric consensus phylogeny of placental mammal species that define divergence nodes which are postulated to have occurred during the mammalian radiations that led to *Homo sapiens*. The high quality genome sequence of each outgroup species would provide a critical step-wise context for the "reconstruction" of the ancestral mammalian genome, which has in time led to our own.