Sequencing and Annotating New Mammalian Y Chromosomes
A White Paper Proposal, July, 2006

Steve Rozen[1], Wesley C. Warren[2], George Weinstock[3], Stephen J. O'Brien[4], Richard A. Gibbs[3], Richard K. Wilson[2], David C. Page[1]

[1] Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142
[2] Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108
[3] Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030
[4] Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland 21702

Overview: Y chromosome sequence is essential for understanding human and mammalian evolution and biology

The only Y chromosome that has been completely sequenced in any species is that of humans[1,2], although sequencing of the chimpanzee and mouse Y chromosomes is in progress[2-4]. Here we propose to sequence and annotate the Y chromosomes of seven additional mammals at varying distances from human: *Macaca mulatta* (rhesus macaque) *Callithrix jacchus* (white-tufted-ear marmoset), *Rattus norvegicus* (laboratory rat), *Bos taurus* (domestic bull), *Felis catus* (domestic cat), *Canis familiaris* (dog), and *Monodelphis domestica* (laboratory opossum). We recognize that NHGRI approved rat Y chromosome sequencing as part of additional rat finishing[5] and that the bull Y chromosome falls within the scope of the Bovine Genome Project, led by the Baylor sequencing center[6]. Inclusion of the rat and bull Y chromosomes is intended (i) to clarify collaborative intentions among the submitting sequencing centers (ii) to ensure coordination and agreement of the definitions of the most useful levels of completion of these chromosomes and (iii) to emphasize coherent biological rationales for sequencing and annotating the seven proposed Y chromosomes in terms of their relevance to human evolution, biology, and disease.

Why sequence and annotate additional mammalian Y chromosomes, and why these species?

- For each of these mammals, NHGRI has the goal of obtaining high-quality genome sequence. However, existing and planned assemblies do not include these Y chromosomes[7-10], which therefore require targeted efforts.
- The human and mouse Y chromosomes are highly enriched in genes essential for spermatogenesis, and available data indicate that this is true for other mammals as well[1,11-17]. Y chromosomal genes are also implicated in cancer, Turner syndrome, graft rejection and graft-versus-host disease, locomotion, high blood pressure, and stress response[1,18-33]. Therefore, biological study of the proposed species requires the annotated sequences of their Y chromosomes.
- Six of the proposed mammals are important laboratory organisms[34-40], and the seventh is an intensively studied domestic animal[6]. The proposed Y chromosome sequences and annotation will support the widespread investigation carried out in these mammals. This, in turn, will advance our understanding of mammalian and human biology, including notably reproductive biology.
- These seven species stand at a variety of evolutionary distances from human, distances chosen to support interpretation of the human Y sequence and its evolution[41,42]. The opossum will provide a critical outgroup for the placental (eutherian) mammals,

including humans. In addition, sequence of the rat Y chromosome will inform our understanding of Y-chromosome sequence in mouse, the preeminent model for mammalian genetics. This will enable better connections to be drawn between mouse and human genetics and reproductive biology.

- 40 years ago, conventional wisdom held that the human Y chromosome had the sole function of triggering testis development[43-45]. Recently, however, genomic studies have revealed unexpected functions and biology. For example, we now know that human Y chromosomes contain numerous genes involved in spermatogenesis, and that interstitial deletions of the Y chromosome are among the most common known causes of human spermatogenic failure[11-14]. We also know that most testis-specific genes in the human Y are present in multiple, highly similar copies, often organized as large, palindromic inverted repeats subject to frequent gene conversion, which may be related to their evolutionary survival or cellular functions[46]. We expect that the sequencing and annotation of the proposed Y chromosomes will lead to further unexpected insights into mammalian and human biology.

Thus, as presented in detail below, the annotated Y-chromosome sequences of these widely studied mammals are essential for understanding key aspects of mammalian biology, including reproductive biology, and for maximizing information about human biology that can be learned from experimental study of these organisms.

Contents

A. Specific biological rationales for the utility of new sequence data

We propose explicit criteria by which additional mammalian Y chromosomes should be prioritized for sequencing at this time:

1. There should be a commitment to sequence the rest of the genome to a "deep draft" (generally defined as at least a 6X coverage).

2. The sequence will inform our understanding of human sequence and human and mammalian biology, for example, by representing one of a range of evolutionary distances from human.

3. The species should be already utilized in active research – including experimental research, since the support for experimental researchers in terms of Y chromosome sequence has lagged behind other genome efforts.

As detailed below, the seven species proposed here score highly in all three criteria. In the future, as Y chromosome sequencing becomes more efficient, it will be feasible to include other species based primarily on criteria 1 and 2. Currently, however, the needs of experimental researchers are underserved by available Y chromosome sequence, and criterion 3 therefore can be weighted more heavily.

A.1/A.2. Improving human health and informing human biology

The sequencing and annotation proposed here will improve human health and inform human biology in three ways.

1. *By fostering and supporting experimental studies of mammalian biology in non-human species* (Sections A.5 and A.6/A.7, below). These studies will advance human reproductive biology and reproductive medicine, as well as other aspects of mammalian and human biology, including germ cell cancers, Turner syndrome, graft-host interactions, locomotion, hypertension, and stress response[1,18-33].
2. *By informing our understanding of the functions of human sequence and by making better connections between human and non-human sequence* (Section A.3/A.4, below). Availability of the proposed sequences will further support experimental studies in the seven proposed mammals, and thereby inform our understanding of the functions of related human sequences. In addition, comparative analysis will identify sequence elements that are conserved between humans and the mammals proposed here and which may be functional. This, in turn, will suggest new studies in the seven mammals.
3. *By improving our understanding of the evolutionary and selective forces operating on mature Y chromosomes, such as those in humans* (Section A8, below). Elucidation of connections between the evolution of mammalian Y chromosomes and their gene content has led to strides in understanding Y-linked clinical phenotypes[1,18-33,47]. Nevertheless, many questions remain. Comparisons of sequence, gene content, and genomic structures between the proposed Y chromosomes and those of human, chimpanzee, and mouse will provide much-needed data for interpreting Y chromosome evolution and, by extension, Y-linked clinical phenotypes[3,17,42,48-54].

In short, understanding the Y chromosomes of these widely studied experimental mammals is essential to understanding key aspects of human biology.

A.3/A.4. Informing the human sequence and providing better connections between human and non-human sequences

The annotated sequences of the proposed Y chromosomes will enable improved interpretation of human sequence, as follows.

1. Alignment of Y chromosomal sequences from each of the species will reveal conserved sequences that represent likely functional elements, both coding and non-coding. Human/rhesus/marmoset comparisons will enable identification of primate-specific conserved elements, while opossum sequences will enable identification of ancient conserved elements[41]. Currently, there is a wealth of comparative sequence data for all human chromosomes *except the Y*. Mammalian genomes that are or will be available at high coverage for *all but the Y-chromosome* include: orangutan, gibbon, rhesus, marmoset, tree shrew, rat, guinea pig, rabbit, bat, cow, dog, cat, armadillo, elephant, and opossum[9,10]. In contrast, comparative data to support interpretation of the human Y chromosome is practically absent.
2. "Informing the human sequence" and "providing better connections between human and non-human sequences" will entail understanding the sequence and functions of the human Y chromosome in the evolutionary context of multiple mammalian Y chromosomes. The three mammalian Y chromosomes currently sequenced or being sequenced (human, chimpanzee, and mouse) are valuable, but do not indicate which aspects of the human Y chromosome are typical or atypical among mammals. Open questions include the following: Are the abundant non-coding but spliced and polyadenylated transcripts observed in the human Y found in other mammalian Y chromosomes, and are they functional? How often have testis-gene-bearing palindromes arisen in mammalian Y chromosome evolution? What is the chromosome biology associated with these genomic structures, especially in germ cells?
3. Over evolution, Y chromosomes have collected spermatogenesis genes, and the Y chromosomes of different mammalian groups acquired or elaborated different sets of spermatogenesis genes[15-17,55-57]. Thus, study of the Y chromosomes of the proposed seven species will guide researchers to spermatogenesis factors that are *not* Y-linked in humans but nonetheless may be important in human reproductive biology.

A.5. Expanding our understanding of biological processes relevant to human health

Most predictably, annotated sequences of the proposed Y chromosomes will expand our understanding of biological processes because these chromosomes and their genes can be studied experimentally. Section A.6/A.7 details expected research opportunities, which involve the following biological processes:

1. *Sex determination and spermatogenesis*. It has been known for 50 years that eutherian Y chromosomes trigger embryonic testis development. However, the sequence of the opossum Y should shed new light on mammalian sex determination and differentiation, since the testis-determining gene shows much more widespread embryonic expression in marsupials than in human or mouse[58]. Beyond sex determination, human and mouse Y chromosomes are essential for spermatogenesis and fertilization. In men seeking

infertility treatment, partial deletions of the Y chromosome are the leading known cause of low or absent sperm count, accounting for ~15% of cases of nonobstructive azoospermia (no sperm in semen) or very low sperm counts ($<5x10^6$/ml)[11-14,59,60]. In mice, some Y-chromosomal genes are needed for proliferation of spermatogonial stem cells, while others are needed for normal sperm morphology and for preventing an excess of female offspring[15,16,61-63]. The proposed sequences will shed additional light on the genetics of mammalian spermatogenesis. For example, the cat Y sequence may shed light on the common defects in sperm morphology, motility, and fertilization that have been extensively studied in this species[64-68].

2. *Germ cell tumors.* Women with ovarian tissue but Y chromosomes are likely to develop gondadoblastoma, a germ cell tumor[18,21,22,24]. The Y-chromosomal gene or genes responsible have not been identified. Men whose Y chromosomes lack a particular 1.6-Mb segment have elevated risks of both low sperm count and testicular germ cell tumors[32,60].
3. *Turner syndrome.* Turner syndrome is caused by a 45,X0 karyotype. Although the Turner phenotype derives partly from haploinsufficiency of the pseudoautosomal *SHOX* gene[69], evidence points to the involvement of genes in the male-specific region of the Y chromosome as well[19]. Many of these genes are absent from the mouse Y chromosome but present in the proposed Y chromosomes, where they can be studied.
4. *Locomotion, high blood pressure, and stress response. Sry* triggers embryonic testis development, but, in the mouse, *Sry* is also expressed in the substantia nigra of the adult brain. There, reduced levels of SRY protein cause reduced tyrosine hydroxylase levels, and, consequently, defects in locomotion[33]. SRY also regulates tyrosine hydroxylase in the rat brain[70], and appears to partly account for the phenotype of the Spontaneously Hypertensive Rat (SHR), which is genetically predisposed to high blood pressure[20,25,27]. Does the male-specific *Sry* gene have nervous system functions in other mammals? Are other non-testicular phenotypes linked to the Y chromosomes of other mammals? The proposed sequencing will help answer questions such as these.

A.6/A.7. Providing additional surrogate systems for human experimentation and facilitating the ability to do experiments and interpret experimental results

A key criterion in proposing these seven mammals is their widespread use in experimental studies, which will be advanced by annotated Y-chromosomal sequences. These species are used as surrogates for humans in studies of germ cell development and fertilization, biological processes that are heavily influenced by Y chromosomal genes. For example, sequences of the rhesus and marmoset Y chromosomes will support this research in model systems that are much more similar to humans than are mice. With rhesus and marmoset it would be possible to study the effects of primate Y-chromosomal genes (genes shared by the human Y chromosome) on early embryonic events, including sex determination, germ cell sex-determination, proliferation of spermatogonial stem cells, meiosis, and spermiogenesis—maturation of haploid meiotic products into spermatozoa. These studies could also encompass investigation into the chromatin structures, epigenetics, and transcriptional activities of Y (and X) chromosomes in spermatogonia, through meiosis, and during spermiogenesis. Rapid advances in RNAi, spermatogonial stem cell culture, transgenics based on spermatogonial stem cells, and culturing of explanted gonads will

facilitate the experimental investigation of Y-chromosomal genes in these seven mammals in the near future[71-75].

A.8. Expanding our understanding of evolutionary processes in general, and of human evolution in particular

Study of Y-chromosome evolution has sought to support and refine the model that X and Y chromosomes have repeatedly – in multiple independent groups of plants and animals – descended from ordinary pairs of autosomes[44,76]. Most prominently, the trajectory and causes of Y-chromosome degeneration have been studied theoretically and empirically in evolutionarily young Y chromosomes that have decayed rapidly[77-82]. Until recently however, almost no data have illuminated the evolutionary processes operating on mature Y chromosomes, such as those in men and other mammals. Three years ago, the only sequenced Y chromosome, that of humans, was found to be a mosaic of several different types of sequence, each reflecting different selective forces and evolutionary events[1]. The "X-degenerate" sequences consist of single copy genes that were formerly allelic to X-chromosomal genes. These sequences best approximate a model of Y evolution as ineluctable degeneration. The "ampliconic" sequences consist of segmental duplications, some extremely large and nearly identical. They harbor numerous, multi-copy genes and transcripts with testis-specific expression. Ampliconic sequences had diverse origins, with some arising from X-degenerate sequences and others from transposed or retrotransposed autosomal sequences.

Comparative approaches to the study of mammalian Y chromosomes will offer new insights into their evolution and functional constraints. For example, a recent comparison of human and chimpanzee X-degenerate sequence found degeneration in the chimpanzee but not human lineage[3]. This study hinted at strong but episodic genetic hitchhiking[77] as a factor in the degeneration of mature Y chromosomes. It was also recently reported that the cat has acquired or elaborated testis-specific genes not seen in human, mouse or chimpanzee[17]. However, much remains to be learned. For example, are the human X-degenerate genes simply a random collection of accidental remnants? Or did some genes persist because they were more important than others? Such questions cannot be addressed from a human-centric point of view, which cannot identify Y-chromosomal genes lost in the human lineage but retained in others. As another example, in human evolution, genes in the recombining tips of the short arms of X and Y chromosomes (the Yp pseudoautosomal region) became recombinationally isolated from each other in stepwise fashion[83-85]. The sequences proposed here will show how recombinational isolation and degeneration proceeded in other mammalian lineages.

Additional important questions revolve around the ampliconic regions of the human Y chromosome. Even the chimpanzee Y chromosome has ampliconic characteristics unlike the human Y chromosome, and mouse ampliconic sequence is radically different (our unpublished analysis). Y chromosomes of tammar wallaby, dog, cat, bull, rat, and numerous apes and monkeys also have ampliconic sequences[17,86-91]. How do ampliconic sequences differ among mammalian Y chromosomes, and do the differences reflect selective forces or random accident?

The opossum Y chromosome will be an important outgroup for analysis of the evolution of human and eutherian Y chromosomes. For example, a large autosomal

segment was translocated to the X and Y chromosomes in a eutherian ancestor *after* divergence from marsupials[92,93]. Thus, the sequence of the opossum Y, which never acquired this translocation, can illuminate the long-term consequences of the eutherian translocation. As another example, Y degeneration likely drove the spread of X inactivation in eutherian evolution[94]. Did analogous processes occur in marsupials, where the paternal X chromosome is always inactivated? Did multi-copy testis-gene families (besides *RBMY* [ref. [88]]) arise in the opossum lineage as they did in eutherians?

B. Strategic issues in acquiring new sequence data

B.1. Size of the research community / demand for sequence

There are large research communities studying the proposed species, which was a major reason that they were selected by NHGRI for high-quality genomic sequencing[6,8,34-40,95,96]. The Y chromosomes of rhesus and marmoset are important because these species are (i) similar to human, (ii) much better models of human spermatogenesis than mouse, and (iii) widely used in biomedical research[97]. The opossum is critical as a non-eutherian outgroup that is easily bred in captivity and that is especially suited to embryological and developmental studies because birth occurs at an early embryonic stage and because of favorable embryonic geometry[40,95]. Dog and cat, due to their popularity as companion animals, enjoy more medical surveillance and literature than any mammalian species other than human. The numbers of PubMed citations and reproductive-medicine-related citations for the seven species proposed here attest to their importance in research (**Table 1**).

Table 1 PubMed citations for the proposed species.

Species	Total Number of PubMed Citations	Number of Reproductive-Biology[a] PubMed Citations
Rhesus	33,197	4,866
Marmoset [b]	2,323	415
Rat	630,527	90,251
Bull/Cow	292,637	33,614
Cat	136,606	2,613
Dog	268,837	6,502
Opossum [b]	466	89

[a] PubMed query "*species* AND (sperm OR spermatogenesis OR testis OR testes OR fertilization OR reproduction)".

[b] Includes entire genus.

Letters of support are provided in the Appendix.

B.2. Suitability of target species for experimentation

The proposed mammals are routinely bred in captivity and widely used in experimental research (**Table 1**, refs. [6,34-40]). These are central reasons that NHGRI selected the proposed species for high-quality sequencing. In particular, rhesus, rat, cat, and dog are mainstays of biomedical research. The opossum, in addition to representing an important outgroup for interpreting eutherian spermatogenesis, is well suited to experimental study of

the early phases of sex differentiation, which occur primarily after birth in this species[98]. Furthermore, even prenatal stages are easier to study in the opossum than in the mouse[40,95]. Spermatogenesis in the bull is intensively studied, and many assisted reproduction techniques now used in human infertility treatment were developed in cattle.

B.3. The rationale for finished Y sequence in the target species

The only sequencing approach that has been successful for any Y chromosome is one based on the mapping and sequencing of a tiling path of large insert clones, such as BACs. This approach was essential for disentangling the large, nearly identical, gene-rich segmental duplications that abound in the human, mouse and chimpanzee Y chromosomes (refs. [1,13] and our unpublished analysis). All experience and theory indicate that approaches based on small insert clones, or even cosmids or fosmids, cannot tease apart different copies of large (often >200 kb) segmental duplications. The proposed Y chromosomes, like those already sequenced or being sequenced, abound in biologically important, gene rich, segmentally duplicated sequence[17,86-91,99]. Thus, they, too, must be sequenced using a BAC-based approach if the resulting sequence is to advance study of the critical questions about Y chromosome biology and evolution discussed above.

B.4. Sequencing strategy, cost, and readiness

Sequencing Strategy

Given the necessity of a BAC-based sequencing approach, we propose to adapt strategies that were proven in the sequencing of the Y chromosomes of human, chimpanzee, and mouse[1,3,13,100]. Here, we propose to aggressively pursue efficiency improvements, including:

- Increased efficiency of the mapping and sequencing of Y chromosomes through greater uniformity and automation of mapping, sequencing, and annotation pipelines. In the annotation pipeline we propose cDNA selection and 454 sequencing of normalized cDNA from several tissues to provide empirical data for gene prediction, especially when ESTs resources are insufficient (**Table 2**, refs. [17,47,101]).
- Reliance on explicit, biologically justified trade offs between the quality and completion of the Y sequence versus cost. Specifically, we propose:
    - For a substantial fraction of the BACs, sequencing will stop at the shotgun or pre-finish phases. Our experience indicates that for many segmentally duplicated, ampliconic regions, prefinish data suffice to disentangle different repeat copies. Regions of particular interest can be finished when necessary.
    - For some large, highly repetitive Y chromosomes, adequate biological knowledge can be obtained by capturing all unique sequences and exemplars of every segmental duplication (ampliconic repeat copy), together with an estimate of their copy number, gene content, and heterogeneity.

In addition, we propose that:

- Representation of all sequences and sequence families in the chromosome is essential. Without this, it is impossible to draw unbiased conclusions about mammalian Y chromosome evolution, and biologically important genes will be missed.

- It is critical to lay the groundwork for potential future, higher quality and lower cost sequencing. For the current phase we regard deep BAC coverage of the Y as an essential component. Where clone coverage is not currently available from existing libraries, we will generate new clone resources. The cost of the additional BAC libraries is minimal, and as an additional cost-saving measure, we propose just-in-time arraying ($~5,000 per library, and $~9,200 per each 1X of arrayed clones – see Appendix). Thus, for species in which we primarily focus on a single exemplar of each sequence, the additional depth will have minimal immediate cost, but library resources will be ready for future refinement as needed.
- The Y chromosomes of several of the mammals proposed here should be sequenced to high quality in their entirety, with priority assigned based on (i) information provided for interpreting human and mammalian sequence, biology, and evolution, and (ii) importance to experimental researchers. Final decisions concerning which chromosomes or chromosomal regions are to be sequenced to the highest quality will be made through consultation between the sequencing centers, collaborators, and NHGRI staff. The decisions will weigh all technical issues along with the significance of preliminary biological findings as the work unfolds.

Sequencing and annotation will entail several concurrent activities, to be pursued independently on each of the seven proposed Y chromosomes:

1. *Generate additional male BAC library coverage.* There is ~5X BAC coverage of the haploid (hemizygous) Y chromosomes of the proposed species (**Table 2**). To avoid numerous gaps in BAC tiling paths and excessive overlaps between BACs, we propose to augment this coverage by generating additional male BAC libraries. As discussed above, to save costs we will array the libraries as needed.

Table 2 Resources available for the seven proposed mammalian species.

	Macaque	Marmoset	Rat	Bull	Cat	Dog	Opossum
♂ **BAC libraries (autosomal/ diploid coverage)**	CHORI-250 (11X)	CHORI-259 (12X)	RPCI-32 (11X)	CHORI-240, RPCI-42, TAMBT, RZPD-754, RZPD 750	RPCI-86 (10X)	RPCI-81 (8X)	VMRC-6 (10X)
♂ **BAC fingerprint map?**	10X CHORI-250	Planned ~10X CHORI-259	No	14X from 3 libraries; 9X from CHORI-240	No	No	No
♂ **BAC end sequences?**	CHORI-250	Planned, CHORI-259	No	CHORI-240 (ref [102])	~3,840	No	No
♂ **Radiation hybrid panel?**	Yes[103]	Requested here	No[a]	Yes[104-106]	Yes[49,107,108]	Yes[109,110]	Requested here
# **ESTs (X10^3)**[b]	55	0	871	1,039	919	359	200 planned[40]

[a] Panel used in refs. [111-113] is from a female.

[b] From http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

2. *Develop Y overgo probes and STSs.* Flow-sort Y chromosomes. Sequence small-insert libraries constructed from the flow-sorted DNA or 454 sequence it. Design Y-chromosome overgo probes (for BAC identification by hybridization) and STSs. This will ensure essentially complete representation of Y chromosomal sequences.

3. *Identification and preliminary characterization of candidate BACs.* Using filter hybridization with Y-chromosome overgo probes, identify BACs representing the target chromosome. Carry out an initial characterization of their overlap by BAC-end sequencing, restriction fingerprinting and STS-content analysis.

4. *Initial BAC selection and sequencing.* Select BACs to form hypothetical contigs and tiling paths. ("Hypothetical" because many apparent BAC overlaps will eventually be revealed as pseudo-overlaps between BACs that represent different copies of a particular segmental duplication.) BACs in hypothetical tiling paths will be subcloned, and shotgun sequenced, using standard procedures. Selected BACs will be finished, depending on biological importance, while others will be sequenced to the prefinish stage.

5. *Distinguishing repeat copies and finding true tiling paths.* For those Y chromosomes where sequencing of each copy of multiple repeats is biologically justified, disentangle repeat copies as follows. In each hypothetical tiling path:

   5.1. Examine BAC overlaps for sequence differences ("sequence family variants", or "paralogous variants") that distinguish between different copies of a sequence family[13,114].

   5.2. PCR amplify and sequence these variants in multiple BACs (based on their map position) to identify true overlaps.

   5.3. Select new tiling paths in which overlaps are consistent with this new information, and then sequence the BACs newly selected at this step.

   Iterate the steps above as necessary until all copies of the sequence family have been identified and all overlaps represent single genomic segments.

6. *Extension and joining of BAC contigs.* Identify clones that extend outward from or link existing BAC contigs. This activity may involve additional BAC identification by filter hybridization and may use sequence family variants, as in the preceding step, to confirm that a clone used to extend a map contig is from the correct repeat copy.

7. *BAC contig ordering, orientation, and gap sizing.* Use FISH and/or radiation hybrid mapping to determine the order and orientation of BAC contigs. This information will allow efficient gap closure, assessment of overall progress, and will provide a final check on large-scale sequence organization when warranted. Extended chromatin FISH will be used to size gaps.

8. *Annotation I: Gene and transcript prediction and testing.* For species for which EST resources are absent or insufficient (**Table 2**), we will use cDNA selection and/or 454 sequencing to provide empirical data on transcription [101]. We will electronically predict transcripts using ESTs and/or other data and then experimentally validate and

determine transcript sequence by techniques including RT-PCR, Northern blotting, and cDNA clone sequencing. Such experimental validation is important because:

8.1. On the human and chimpanzee Y chromosomes, many electronically predicted genes proved to be false positives (refs. [2,3] and unpublished observations), and this will likely be the case for the proposed Y chromosomes.

8.2. Because mammalian Y chromosomes have undergone rapid evolution, genes on the proposed chromosomes may not share expression and splicing patterns with their orthologs in human, chimpanzee, or mouse[2,3].

8.3. Because of differential preservation of X-degenerate genes and differential acquisition and elaboration of multi-copy testis genes, the proposed Y chromosomes may have genes not present in the human, chimpanzee, or mouse Y chromosomes.

9. *Annotation II: Comparative and evolutionary analysis.* Perform comparative and evolutionary analyses on the sequenced Y chromosomes, including:

9.1. Comparison of ampliconic (segmental duplication) organization between the proposed Y chromosomes and those of human, chimpanzee, and mouse. Which ampliconic genes are common among multiple species? Which were elaborated independently? How do their organizations compare between species? What are the evolutionary origins of ampliconic sequences? What homologous genes in humans and other organisms might have related functions?

9.2. Comparative sequence analysis. What X-degenerate genes are present in other mammals but not in human, mouse, or chimpanzee? What genes seem to be conserved based on analyses such K_a/K_s? Are there conserved, non-coding elements?

Sequencing Cost

We estimate total sequencing and annotation costs to be supplied through already allocated sequencing funds at $~8,000,000.

Sequencing Readiness (and Readiness for Analysis and Annotation)

The proposed species are ready for Y-chromosome sequencing and annotation and satisfy two critical prerequisites:

- Availability of high quality sequence from the rest of the genome, at a minimum 6X draft (refs. [8,9,96], and http://www.genome.gov/12512284, http://www.genome.gov/12512285). This is essential for comparing Y chromosomal genes to homologous genes on the X chromosome and autosomes.
- Availability of male BAC libraries with hybridization filters.

Additional resources are available for some species (**Table 2**). These include:

- BAC-end sequences and/or BAC fingerprint maps from male BAC libraries, with which to select BACs for tiling paths,
- Male radiation-hybrid mapping panels, with which to order and orient BAC contigs.

- ESTs to provide empirical support for gene prediction.

For the rat, ~2X of short-insert reads are available from flow sorted Y chromosomes[5] and will be used to design hybridization probes and STSs for BAC selection and mapping.

B.5. Other (partial) sources of funding

The laboratory of David C. Page expects to support its activities under this proposal (mapping, annotation, and analysis) with ~$5,000,000 (direct and indirect) from NIH/NHGRI grant HG000257 and Howard Hughes Medical Institute.

References

1. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-837 (2003).
2. Hughes, J.F., Skaletsky, H., Rozen, S., Wilson, R.K. & Page, D.C. Has the chimpanzee Y chromosome been sequenced? *Nature Genet.*, In press (2006).
3. Hughes, J.F. et al. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437**, 100-103 (2005).
4. Kuroki, Y. et al. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nature Genet.* **38**, 158-167 (2006).
5. Gibbs, R.A. & Weinstock, G.M. Upgrading the DNA Sequence of the Rat Genome, white paper available at http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/RatUpgradeSeq.pdf. (2005).
6. Gibbs, R.A. et al. Bovine genome sequencing initiative, cattle-izing the human genome (white paper proposing sequencing of the bovine genome), available at http://www.genome.gov/12512284. (2002).
7. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
8. Gibbs, R.A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).
9. National Human Genome Research Institute. Previous Genomic Sequencing Targets (June, 2006) http://www.genome.gov/19516773. (2006).
10. National Human Genome Research Institute. New Genomic Sequencing Targets http://www.genome.gov/11007951. (2006).
11. Reijo, R. et al. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nature Genet.* **10**, 383-393 (1995).
12. Vogt, P.H. et al. Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. *Hum. Mol. Genet.* **5**, 933-943 (1996).
13. Kuroda-Kawaguchi, T. et al. The *AZFc* region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279-286 (2001).
14. Repping, S. et al. Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am. J. Hum. Genet.* **71**, 906-922 (2002).
15. Ellis, P.J. et al. Deletions on mouse Yq lead to upregulation of multiple X- and Y-linked transcripts in spermatids. *Hum. Mol. Genet.* **14**, 2705-2715 (2005).
16. Toure, A. et al. Identification of novel Y chromosome encoded transcripts by testis transcriptome analysis of mice with deletions of the Y chromosome long arm. *Genome Biology* **6**, R102 (2005).
17. Murphy, W.J. et al. Novel gene acquisition on carnivore Y chromosomes. *PLoS Genet* **2**, e43 (2006).

18. Page, D.C. Hypothesis: A Y-chromosomal gene causes gonadoblastoma in dysgenetic gonads. *Development* **101 Suppl.**, 151-155 (1987).
19. Fisher, E.M. et al. Homologous ribosomal protein genes on the human X and Y chromosomes: Escape from X inactivation and possible implications for Turner syndrome. *Cell* **63**, 1205-1218 (1990).
20. Ely, D.L., Daneshvar, H., Turner, M.E., Johnson, M.L. & Salisbury, R.L. The hypertensive Y chromosome elevates blood pressure in F11 normotensive rats. *Hypertension* **21**, 1071-1075 (1993).
21. Page, D.C. Y chromosome sequences in Turner's syndrome and risk of gonadoblastoma or virilisation [letter; comment]. *Lancet* **343**, 240 (1994).
22. Salo, P. et al. Molecular mapping of the putative gonadoblastoma locus on the Y chromosome. *Genes. Chromosomes Cancer* **14**, 210-214 (1995).
23. Wang, W. et al. Human H-Y: A male-specific histocompatibility antigen derived from the SMCY protein. *Science* **269**, 1588-1590 (1995).
24. Tsuchiya, K., Reijo, R., Page, D.C. & Disteche, C.M. Gonadoblastoma: molecular definition of the susceptibility region on the Y chromosome. *Am. J. Hum. Genet.* **57**, 1400-1407 (1995).
25. Ely, D. et al. Spontaneously hypertensive rat Y chromosome increases indexes of sympathetic nervous system activity. *Hypertension* **29**, 613-618 (1997).
26. Pierce, R.A. et al. Cutting edge: the HLA-A*0101-restricted HY minor histocompatibility antigen originates from *DFFRY* and contains a cysteinylated cysteine residue as identified by a novel mass spectrometric technique. *J. Immunol.* **163**, 6360-6364 (1999).
27. Dumas, P. et al. Contribution of autosomal loci and the Y chromosome to the stress response in rats. *Hypertension* **35**, 568-573 (2000).
28. Vogt, M.H. et al. *UTY* gene codes for an HLA-B60-restricted human male-specific minor histocompatibility antigen involved in stem cell graft rejection: characterization of the critical polymorphic amino acid residues for T-cell recognition. *Blood* **96**, 3126-3132 (2000).
29. Warren, E. et al. The human *UTY* gene encodes a novel HLA-B8-restricted H-Y antigen. *J. Immunol.* **164**, 2807-2814 (2000).
30. Vogt, M.H. et al. The *DBY* gene codes for an HLA-DQ5-restricted human male-specific minor histocompatibility antigen involved in graft-versus-host disease. *Blood* **99**, 3027-3032 (2002).
31. Spierings, E. et al. Identification of HLA class II-restricted H-Y-specific T-helper epitope evoking CD4+ T-helper cells in H-Y-mismatched transplantation. *Lancet* **362**, 610-615 (2003).
32. Nathanson, K.L. et al. The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *Am. J. Hum. Genet.* **77**, 1034-1043 (2005).
33. Dewing, P. et al. Direct regulation of adult brain function by the male-specific factor SRY. *Curr. Biol.* **16**, 415-420 (2006).
34. O'Brien, S.J. et al. Sequencing the Genome of the Domestic Cat *Felis catus*, white paper available at http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/CatSEQ.pdf. (2002).

35. O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J. & Yuhki, N. The Feline Genome Project. *Ann. Rev. Genet.* **36**, 657-686 (2002).
36. Ostrander, E.A., Lindblad-Toh, K. & Lander, E.S. Sequencing the Genome of the Domestic Dog *Canis familiaris*, white paper available at http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/CanineSEQedited.pdf. (2002).
37. Rogers, J., Katze, M., Bumgarner, R., Gibbs, R.A. & Weinstock, G.M. White Paper for Complete Sequencing of the Rhesus Macaque (*Macaca mulatta*) Genome, available at http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/RhesusMacaqueSEQ021203.pdf. (2002).
38. Amemiya, C.T. et al. Proposal for Complete Sequencing of the Genome of a Marsupial: The Gray, Short-tailed Opossum, *Monodelphis domestica*, white paper available at http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/OpossumSEQ.pdf. (2003).
39. Mansfield, K., Tardif, S. & Eicher, E. White Paper for Complete Sequencing of the Common Marmoset (*Callithirx jacchus*) Genome, available at http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/MarmosetSeq.pdf. (2004).
40. Samollow, P.B. Status and applications of genomic resources for the gray, short-tailed opossum, *Monodelphis domestica*, and American marsupial model for comparative biology. *Aust. J. Zool.* **54**, 173-196 (2006).
41. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391-1394 (2003).
42. Murphy, W.J., Pevzner, P.A. & O'Brien, S.J. Mammalian phylogenomics comes of age. *Trends Genet.* **20**, 631-639 (2004).
43. Stern, C. The problem of complete Y-linkage in men. *Am. J. Hum. Genet.* **9**, 147-166 (1957).
44. Ohno, S. *Sex Chromosomes and Sex-linked Genes*, (Springer-Verlag, Berlin, 1967).
45. Page, D.C. 2003 Curt Stern Award address. On low expectation exceeded; or, the genomic salvation of the Y chromosome. *Am. J. Hum. Genet.* **74**, 399-402 (2004).
46. Rozen, S. et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873-876 (2003).
47. Lahn, B.T. & Page, D.C. Functional coherence of the human Y chromosome. *Science* **278**, 675-680 (1997).
48. Pecon Slattery, J. & O'Brien, S.J. Patterns of Y and X chromosome DNA sequence divergence during the Felidae radiation. *Genetics* **148**, 1245-1255 (1998).
49. Murphy, W.J., Sun, S., Chen, Z.Q., Pecon-Slattery, J. & O'Brien, S.J. Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Res.* **9**, 1223-1230 (1999).
50. Pecon Slattery, J., Sanner-Wachter, L. & O'Brien, S.J. Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in Felidae (Mammalia). *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5307-5312 (2000).

51. Murphy, W.J., Stanyon, R. & O'Brien, S.J. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biology* **2**, REVIEWS0005 (2001).
52. Murphy, W.J., Bourque, G., Tesler, G., Pevzner, P. & O'Brien, S.J. Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Hum Genomics* **1**, 30-40 (2003).
53. Pecon-Slattery, J., Pearks Wilkerson, A.J., Murphy, W.J. & O'Brien, S.J. Phylogenetic assessment of introns and SINEs within the Y chromosome using the cat family felidae as a species tree. *Mol. Biol. Evol.* **21**, 2299-2309 (2004).
54. Murphy, W.J. et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**, 613-617 (2005).
55. Mazeyrat, S. & Mitchell, M.J. Rodent Y chromosome *TSPY* gene is functional in rat and non-functional in mouse. *Hum. Mol. Genet.* **7**, 557-62 (1998).
56. Mazeyrat, S. et al. A Y-encoded subunit of the translation initiation factor Eif2 is essential for mouse spermatogenesis. *Nature Genet.* **29**, 49-53 (2001).
57. Toure, A. et al. A protein encoded by a member of the multicopy *Ssty* gene family located on the long arm of the mouse Y chromosome is expressed during sperm development. *Genomics* **83**, 140-147 (2004).
58. Harry, J.L., Koopman, P., Brennan, F.E., Graves, J.A. & Renfree, M.B. Widespread expression of the testis-determining gene SRY in a marsupial. *Nature Genet.* **11**, 347-349 (1995).
59. Oates, R.D., Silber, S., Brown, L.G. & Page, D.C. Clinical characterization of 42 oligospermic or azoospermic men with microdeletion of the *AZFc* region of the Y chromosome, and of 18 children conceived via ICSI. *Hum. Reprod.* **17**, 2813-24. (2002).
60. Repping, S. et al. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nature Genet.* **35**, 247-251 (2003).
61. Styrna, J., Imai, H.T. & Moriwaki, K. An increased level of sperm abnormalities in mice with a partial deletion of the Y chromosome. *Genet. Res.* **57**, 195-9. (1991).
62. Styrna, J., Klag, J. & Moriwaki, K. Influence of partial deletion of the Y chromosome on mouse sperm phenotype. *J. Reprod. Fertil.* **92**, 187-195 (1991).
63. Conway, S.J. et al. *Y353/B*: a candidate multiple-copy spermiogenesis gene on the mouse Y chromosome. *Mamm. Genome* **5**, 203-210 (1994).
64. Howard, J.G., Brown, J.L., Bush, M. & Wildt, D.E. Teratospermic and normospermic domestic cats: ejaculate traits, pituitary-gonadal hormones, and improvement of spermatozoal motility and morphology after swim-up processing. *J. Androl.* **11**, 204-215 (1990).
65. Long, J.A. et al. Sperm capacitation and the acrosome reaction are compromised in teratospermic domestic cats. *Biol. Reprod.* **54**, 638-646 (1996).
66. Franca, L.R. & Godinho, C.L. Testis morphometry, seminiferous epithelium cycle length, and daily sperm production in domestic cats (Felis catus). *Biol. Reprod.* **68**, 1554-1561 (2003).
67. Comizzoli, P., Wildt, D.E. & Pukazhenthi, B.S. Poor Centrosomal Function of Cat Testicular Spermatozoa Impairs Embryo Development In Vitro After Intracytoplasmic Sperm Injection. *Biol. Reprod.* **75**, 252-260 (2006).

68. Pukazhenthi, B.S., Neubauer, K., Jewgenow, K., Howard, J. & Wildt, D.E. The impact and potential etiology of teratospermia in the domestic cat and its wild relatives. *Theriogenology* **66**, 112-121 (2006).
69. Rao, E. et al. Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nature Genet.* **16**, 54-63 (1997).
70. Milsted, A. et al. Regulation of tyrosine hydroxylase gene transcription by *Sry*. *Neurosci. Lett.* **369**, 203-207 (2004).
71. Hamra, F.K. et al. Self renewal, expansion, and transfection of rat spermatogonial stem cells in culture. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 17430-17435 (2005).
72. Aponte, P.M., Soda, T., van de Kant, H.J. & de Rooij, D.G. Basic features of bovine spermatogonial culture and effects of glial cell line-derived neurotrophic factor. *Theriogenology* **65**, 1828-1847 (2006).
73. de Rooij, D.G. Rapid expansion of the spermatogonial stem cell tool box. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7939-7940 (2006).
74. Guan, K. et al. Pluripotency of spermatogonial stem cells from adult mouse testis. *Nature* **440**, 1199-1203 (2006).
75. Kanatsu-Shinohara, M. et al. Production of knockout mice by random or targeted mutagenesis in spermatogonial stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8018-9023 (2006).
76. Muller, H.J. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* **3**, 422-499 (1918).
77. Rice, W.R. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116**, 161-167 (1987).
78. Charlesworth, B. The effect of background selection against deleterious alleles on weakly selected, linked variants. *Genet. Res.* **63**, 213-227 (1994).
79. Rice, W.R. Degeneration of a nonrecombining chromosome. *Science* **263**, 230-232 (1994).
80. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 1563-1572 (2000).
81. Rice, W.R. & Chippindale, A.K. Sexual recombination and the power of natural selection. *Science* **294**, 555-559 (2001).
82. Bachtrog, D. & Charlesworth, B. Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* **416**, 323-326 (2002).
83. Toder, R. et al. Genes located in and near the human pseudoautosomal region are located in the X-Y pairing region in dog and sheep. *Chromosome Res.* **5**, 301-306 (1997).
84. Lahn, B.T. & Page, D.C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964-967 (1999).
85. Ross, M.T. et al. The DNA sequence of the human X chromosome. *Nature* **434**, 325-337 (2005).
86. Jakubiczka, S., Schnieders, F. & Schmidtke, J. A bovine homologue of the human *TSPY* gene. *Genomics* **17**, 732-735 (1993).
87. Nagamine, C.M. The testis-determining gene, *SRY*, exists in multiple copies in Old World rodents. *Genet. Res.* **64**, 151-159 (1994).

88. Delbridge, M.L. et al. A human candidate spermatogenesis gene, *RBM1*, is conserved and amplified on the marsupial Y chromosome. *Nature Genet.* **15**, 131-136 (1997).
89. Glaser, B. et al. Simian Y chromosomes: species-specific rearrangements of *DAZ*, *RBM*, and *TSPY* versus contiguity of PAR and *SRY*. *Mamm. Genome* **9**, 226-231 (1998).
90. Gromoll, J. et al. The Old World monkey *DAZ* (Deleted in AZoospermia) gene yields insights into the evolution of the *DAZ* gene cluster on the human Y chromosome. *Hum. Mol. Genet.* **8**, 2017-2024 (1999).
91. Verkaar, E.L. et al. Organization and concerted evolution of the ampliconic Y-chromosomal *TSPY* genes from cattle. *Genomics* **84**, 468-474 (2004).
92. Watson, J.M., Spencer, J.A., Riggs, A.D. & Graves, J.A. Sex chromosome evolution: Platypus gene mapping suggests that part of the human X chromosome was originally autosomal. *Proc. Natl. Acad. Sci. USA* **88**, 11256-11260 (1991).
93. Graves, J.A.M. The origin and function of the mammalian Y chromosome and Y-borne genes — an evolving understanding. *Bioessays* **17**, 311-321 (1995).
94. Jegalian, K. & Page, D.C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776-780 (1998).
95. Renfree, M.B. & Shaw, G. Germ cells, gonads and sex reversal in marsupials. *Int. J. Dev. Biol.* **45**, 557-567 (2001).
96. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
97. Millar, M.R., Sharpe, R.M., Weinbauer, G.F., Fraser, H.M. & Saunders, P.T. Marmoset spermatogenesis: organizational similarities to the human. *Int. J. Androl.* **23**, 266-277 (2000).
98. Baker, P.J., Moore, H.D., Penfold, L.M., Burgess, A.M. & Mittwoch, U. Gonadal sex differentiation in the neonatal marsupial, *Monodelphis domestica*. *Development* **109**, 699-703 (1990).
99. Kostova, E., Rottger, S., Schempp, W. & Gromoll, J. Identification and characterization of the cynomolgus monkey chromodomain gene *cynCDY*, an orthologue of the human *CDY* gene family. *Mol. Hum. Reprod.* **8**, 702-709 (2002).
100. Page, D.C. & Rozen, S. Genomic Studies of Sex Chromosomes, competing renewal application for NIH/NHGRI HG000257. (2005).
101. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).
102. Larkin, D.M. et al. A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Res.* **13**, 1966-1972 (2003).
103. Murphy, W.J., Page, J.E., Smith, C., Jr., Desrosiers, R.C. & O'Brien, S.J. A radiation hybrid mapping panel for the rhesus macaque. *J. Hered.* **92**, 516-519 (2001).
104. Band, M.R. et al. An ordered comparative map of the cattle and human genomes. *Genome Res.* **10**, 1359-1368 (2000).
105. Liu, W.S., Mariani, P., Beattie, C.W., Alexander, L.J. & Ponce De Leon, F.A. A radiation hybrid map for the bovine Y Chromosome. *Mamm. Genome* **13**, 320-326 (2002).

106. Williams, J.L. et al. A bovine whole-genome radiation hybrid panel and outline map. *Mamm. Genome* **13**, 469-474 (2002).
107. Murphy, W.J., Menotti-Raymond, M., Lyons, L.A., Thompson, M.A. & O'Brien, S.J. Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics* **57**, 1-8 (1999).
108. Murphy, W.J. et al. A radiation hybrid map of the cat genome: implications for comparative mapping. *Genome Res.* **10**, 691-702 (2000).
109. Vignaux, F. et al. Toward a dog radiation hybrid map. *J. Hered.* **90**, 62-67 (1999).
110. Guyon, R. et al. A 1-Mb resolution radiation hybrid map of the canine genome. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5296-5301 (2003).
111. Steen, R.G. et al. A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.* **9**, AP1-8, insert (1999).
112. Watanabe, T.K. et al. A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genet.* **22**, 27-36 (1999).
113. Kwitek, A.E. et al. High-density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence. *Genome Res.* **14**, 750-757 (2004).
114. Saxena, R. et al. Four *DAZ* genes in two clusters found in the *AZFc* region of the human Y chromosome. *Genomics* **67**, 256-267 (2000).