Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species

GENOME 10K COMMUNITY OF SCIENTISTS*

*G10KCOS authors are listed in the Appendix.

Abstract

The human genome project has been recently complemented by whole-genome assessment sequence of 32 mammals and 24 nonmammalian vertebrate species suitable for comparative genomic analyses. Here we anticipate a precipitous drop in costs and increase in sequencing efficiency, with concomitant development of improved annotation technology and, therefore, propose to create a collection of tissue and DNA specimens for 10 000 vertebrate species specifically designated for whole-genome sequencing in the very near future. For this purpose, we, the Genome 10K Community of Scientists (G10KCOS), will assemble and allocate a biospecimen collection of some 16 203 representative vertebrate species spanning evolutionary diversity across living mammals, birds, nonavian reptiles, amphibians, and fishes (ca. 60 000 living species). In this proposal, we present precise counts for these 16 203 individual species with specimens presently tagged and stipulated for DNA sequencing by the G10KCOS. DNA sequencing has ushered in a new era of investigation in the biological sciences, allowing us to embark for the first time on a truly comprehensive study of vertebrate evolution, the results of which will touch nearly every aspect of vertebrate biological enquiry.

Key words: ancestral state reconstruction, comparative genomics, G10K, molecular evolution, species conservation, vertebrate biology

The bold insight behind the success of the human genome project was that, although vast, the roughly 3 billion letters of digital information specifying the total genetic heritage of an individual is finite and might, with dedicated resolve, be brought within the reach of our technology (Lander et al. 2001; Venter et al. 2001; Collins et al. 2003). The number of living species is similarly vast, estimated to be between $10^6$ and $10^9$ for metazoans and approximately $6 \times 10^7$ for Vertebrata, which includes our closest relatives (May 1988; Erwin 1991; Gaston 1991). With the same unity of purpose shown for the Human Genome Project, we can now contemplate reading the genetic heritage of all species, beginning today with the vertebrates. The feasibility of a “Genome 10K” (G10K) project to catalog the genomic diversity of 10 000 vertebrate genomes, approximately one for each vertebrate genus, requires only one more order of magnitude reduction in the cost of DNA sequencing, after the 4 orders of magnitude reduction we have seen in the last 10 years (Benson et al. 2008; Mardis 2008; Shendure and Ji 2008; Eid et al. 2009). The approximate number of 10 000 is a compromise between reasonable expectations for the reach of new sequencing technology over the next few years and adequate coverage of vertebrate species diversity. It is time to prepare for this undertaking.

Living vertebrate species derive from a common ancestor that lived between 500 and 600 million years ago (Ma), before the time of the Cambrian explosion of animal life. Because a core repertoire of about 10 000 genes in a genome of about a billion bases is seen in multiple, deeply branching vertebrates and close deuterostome sister groups, we may surmise that the haploid genome of the common vertebrate ancestor was already highly sophisticated. At a minimum, this genome would have consisted of $10^8$–$10^9$ bases specifying a body plan that included, among other features: 1) segmented muscles derived from somites; 2) a notochord and dorsal hollow neural tube differentiating into primitive forebrain, midbrain, hindbrain, and spinal-chord structures; 3) basic endocrine functions encoded in distant precursors to the thyroid, pancreas, and other vertebrate organs; and 4) a highly sophisticated innate immune system (Aparicio et al. 2002; Dehal et al. 2002; Hillier et al. 2004; Sodergren et al. 2006; Holland et al. 2008; Osorio and Retaux 2008; Gregory 2009). In the descent of the living vertebrates, the roughly $10^9$ bases in the DNA segments that specify these sophisticated features, along with more fundamental biological processes, recorded many billions of fixed changes, the outcome of innumerable natural evolutionary experiments. These and other genetic changes, including rearrangements, duplications, and losses, spawned the diversity of vertebrate forms that inhabit strikingly diverse environments of the planet today. A G10K project explicitly detailing these genetic changes will provide...
an essential reference resource for an emerging new synthesis of molecular, organismic, developmental, and evolutionary biology to explore the vertebrate forms of life, just as the human genome project has provided an essential reference resource for 21st century biomedicine.

Beyond elaborations of ancient biochemical and developmental pathways, vertebrate evolution is characterized by stunning innovations, including adaptive immunity, multichambered hearts, cartilage, bones, and teeth, an internal skeleton that has given rise to the largest aquatic and terrestrial animals on the planet, a variety of sensory modalities that detect and process external stimuli, and specialized endocrine organs such as the pancreas, thyroid, thymus, pituitary, adrenal, and pineal glands (Shimeld and Holland 2000). At the cellular level, the neural crest, sometimes referred to as a fourth germ layer, is unique to vertebrates and gives rise to a great variety of structures, including some skeletal elements, tendons and smooth muscle, neurons and glia of the autonomic nervous system, melanocytes in the skin, dentin in the teeth, parts of endocrine-system organs, and connective tissue in the heart (Meulemans and Bronner-Fraser 2002; Baker 2008). Integration of sophisticated vertebrate sensory, neuroanatomical and behavioral elaborations coupled with often dramatic anatomical and physiological changes allowed exploitation of oceanic, terrestrial, and aerial ecological niches. Anticipated details of expansions and losses of specific gene families revealed by the G10K project will provide new insights into the molecular mechanisms behind these extraordinary innovations.

Adaptive changes in noncoding regulatory DNA also play a fundamental role in vertebrate evolution and understanding these changes represents an even greater challenge for comparative genomics (Hoeckstra and Coyne 2007; Stranger et al. 2007). Almost no part of the known noncoding vertebrate gene regulatory apparatus bears any discernible resemblance at the DNA level to analogous systems in our deuterostome distant cousins. Yet, noncoding DNA segments represent the majority of the bases found to be under selection for the removal of deleterious alleles, and are likely to form the majority of the functional units in vertebrate genomes (Waterston et al. 2002; Siepel et al. 2005). Noncoding DNA segments are also hypothesized to be the major source of evolutionary innovation within vertebrate subclades (King and Wilson 1975; Holland et al. 2008). The origins and evolutionary trajectory of the subset of noncoding functional elements under the strongest selection to remove deleterious alleles can be traced deep into the vertebrate tree (Bejerano et al. 2004), in many cases to its very root, whereas other noncoding functional elements have uniquely arisen at the base of a particular class, order or family of vertebrate species. Within vertebrate lineages that evolved from a common ancestor in the last 100 My, such as placental mammals (~5000 species), modern birds (~10,000 species), and acanthomorphan fishes (~16,000 species), evolutionary coalescence to a common ancestral DNA segment can be reliably determined even for segments of noncoding DNA. This enables detailed studies of base-by-base evolutionary changes throughout the genome, in both coding and noncoding DNA. Thus, the G10K project will provide power to address critical hypotheses concerning the origin and evolution of functional noncoding DNA segments and their role in molding physiological and developmental definitions of living animal species.

Through comprehensive investigation of vertebrate evolution, the G10K project will also lay the foundation needed to understand the genetic basis of recent and rapid adaptive changes within species and between closely related species. Coupled with evolutionary studies of recently diversifying clades, it will help address an increasingly urgent need to predict species’ responses to climate change, pollution, emerging diseases, and invasive competitors (Stockwell et al. 2003; Bell et al. 2004; Kohn et al. 2006; Thomas et al. 2009). It will enable studies of genomic phylogeography and population genetics that are crucial to assessment, monitoring, and management of biological diversity, especially of threatened and endangered species (Brito and Edwards 2009). Recent studies validate some of the potential contributions that the availability of genome sequences can provide to endangered species conservation efforts (Hillier et al. 2004; Romanov et al. 2009). Whole-genome sequence assemblies will be essential to facilitate genome-wide single nucleotide polymorphism discovery and to enable studies of historical demography, population structure, disease risk factors, and a variety of other conservation-related biological attributes. Species for which assembled whole-genome sequences are available will immediately be more amenable to a variety of biological studies that can contribute to assessments and science-based management. Such understanding could help curb the accelerating extinction crisis and slow the loss of biodiversity worldwide. Thus, as many threatened or endangered species should be included in the G10K project as is feasible.

Proposal

To this end, we propose to assemble a “virtual collection” of frozen or otherwise suitably preserved tissues or DNA samples representing on the order of 10,000 extant vertebrate species, including some recently extinct species that are amenable to genomic sequencing (Table 1). This collection represents combined specimen materials from at least 43 participating institutions (Table 2). In many cases, we have collected both male and female samples and for certain species several samples that reflect geographic diversity and/or diversity within localized populations.

Tissues in genetic resource collections are stored by different methods, which yield varying results with regard to DNA quality (Edwards et al. 2005). Tissues that are sampled from the field may be left at ambient temperatures for several hours before they are finally frozen in liquid nitrogen and subsequently stored there at or near ~80 °C. Nonetheless, many of these tissues still yield high-quality DNA (Brumfield R, LSU, personal communication). In other cases, noncryogenic field buffers are used, although
Table 1. Counts of vertebrate species stipulated for Genome 10K collection from G10KCOS

<table>
<thead>
<tr>
<th>Orders</th>
<th>Families</th>
<th>Genera</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>With G10K</td>
<td>Total</td>
<td>% of</td>
<td>With G10K</td>
</tr>
<tr>
<td>with total</td>
<td>samples</td>
<td>total</td>
<td>samples</td>
</tr>
<tr>
<td>Mammals</td>
<td>27</td>
<td>100</td>
<td>145</td>
</tr>
<tr>
<td>Birds</td>
<td>32</td>
<td>94</td>
<td>182</td>
</tr>
<tr>
<td>Amphibians</td>
<td>3</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Reptiles</td>
<td>4</td>
<td>100</td>
<td>63</td>
</tr>
<tr>
<td>Fishes</td>
<td>62</td>
<td>100</td>
<td>424</td>
</tr>
<tr>
<td>Totals</td>
<td>128</td>
<td>98</td>
<td>864</td>
</tr>
</tbody>
</table>


with varying results. In addition to DNA quality, permit and species validation are also important issues to consider (Supplementary Material, Appendix 1). We will follow 4 general guidelines for G10K sample collection:

1. We seek 20 µg of genomic DNA or about 1 g of frozen tissue for each target species.
2. Tissues may be field preserved in liquid nitrogen, ethanol, or dimethyl sulfoxide (DMSO). Initial preservation in liquid nitrogen is strongly recommended for new acquisitions when field conditions permit. Transport from the field in liquid nitrogen or dry ice, or the use of dry shippers at liquid nitrogen temperatures is encouraged. In the laboratory, tissues should be stored in –80 °C freezers or in liquid nitrogen. The “gold standard” that would permit cell lines to be harvested is storage in liquid nitrogen of finely minced tissue fragments equilibrated in tissue-culture medium containing DMSO.
3. Tissues should be documented with voucher specimens linked to institutional accession codes when feasible (preserved carcasses are preferred, although photo vouchers are acceptable) and DNA Barcode information will be collected for all specimens (Hebert et al. 2003; Hanner and Gregory 2007; Ratnasingham and Hebert 2007; Field 2008; Field et al. 2008; Borisenko et al. 2009). In the case of rare or endangered species, a tissue sample, locality, identification by a professional zoologist, and DNA Barcode confirmation or listing in the International Species Information System would be acceptable.
4. All specimens used in the G10K project will be obtained and relocated in accord with national and international statutes regulating the collection, use, and transport of biological specimens.

In addition to samples for DNA extraction, the collection will include 1006 cryopreserved fibroblast cell lines derived from 602 different vertebrate species, primarily mammals, but including representatives of 300 taxa comprising 42 families of nonmammalian amniotes and 1 amphibian species. These resources provide an additional window into the unique cell biology of these species. With the recent development of transformation techniques to create induced pluripotent stem cells from fibroblast lines (Okita et al. 2007; Stadtfeld et al. 2008; Yu 2009; Yusa et al. 2009), the potential of cell-line studies is greatly expanded. Although it is still unclear how well current cell-line generation methods can be extended to all vertebrate clades (Liu et al. 2008; Trounson 2009), we propose to initiate primary fibroblast cell cultures for as many species as possible, with a target of at least 2,000 diverse species, as a corollary outcome of the G10K project. These cell cultures, along with cDNA derived from primary tissues, will provide direct access to gene expression and regulation data in the vertebrate species we catalog and provide a renewable experimental resource to complement the G10K genome sequences. For at least one species of each vertebrate order, we propose to assemble additional genomic resources, including physical maps and a bacterial artificial chromosome (BAC) library, other cell lines, and primary tissues for transcriptome analysis. For these species, we will propose to sequence multiple individuals to assess within-species diversity, including members of both sexes to assess sex-chromosome differences. A resource of this magnitude would help catalyze a much-needed extension of experimental molecular biology beyond the very limited set of model organisms it currently explores.

Integrated analysis and rapid release (genome.gov 2003) of the G10K data represents a substantial informatics challenge, beginning with the construction of a sample tracking database and culminating with the software needed to support a detailed evolutionary analysis of the many terabytes of sequence data (Supplementary Material, Appendices 3 and 4).

The G10K species collection will include tissue/DNA specimens from 5 major organismal groups: mammals, birds, amphibians, nonavian reptiles, and fishes (Table 1, Figure 1). Relevant aspects of each major group compiled by the Taxon committee chairs follow.

**Mammals**

Mammals contain a morphologically and behaviorally diverse assemblage of approximately 5400 species from...
## Table 2. List of collections and participating institutions

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Steward(s)</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Natural Sciences, Philadelphia</td>
<td>Nate Rice</td>
<td><a href="http://www.ansp.org/">http://www.ansp.org/</a></td>
</tr>
<tr>
<td>American Museum of Natural History</td>
<td>Joel Cracraft</td>
<td><a href="http://research.amnh.org/ornithology/personnel/jlc.htm">http://research.amnh.org/ornithology/personnel/jlc.htm</a></td>
</tr>
<tr>
<td>Bell Museum of Natural History, University of Minnesota</td>
<td>F. Keith Barker</td>
<td><a href="http://www.bellmuseum.org/">http://www.bellmuseum.org/</a></td>
</tr>
<tr>
<td>Biodiversity Institute of Ontario</td>
<td>Alexei Bortiseiko</td>
<td><a href="http://www.biodiversity.ca">http://www.biodiversity.ca</a></td>
</tr>
<tr>
<td>Biodiversity Research Institute, University of Kansas</td>
<td>Edward O. Wiley</td>
<td><a href="http://www.nhm.ku.edu/fishes/">http://www.nhm.ku.edu/fishes/</a></td>
</tr>
<tr>
<td>Burke Museum, University of Washington</td>
<td>To be determined</td>
<td><a href="http://www.washington.edu/burkemuseum/collections/genetic/index.php">http://www.washington.edu/burkemuseum/collections/genetic/index.php</a></td>
</tr>
<tr>
<td>California Academy of Sciences</td>
<td>Jens V. Vindum</td>
<td>NA</td>
</tr>
<tr>
<td>CIBIO, University of Porto, Portugal</td>
<td>Albano Beija-Pereira</td>
<td><a href="http://cibio.up.pt/">http://cibio.up.pt/</a></td>
</tr>
<tr>
<td>Círculo Herpetológico de Panamá</td>
<td>Roberto Ibáñez</td>
<td><a href="http://ara.inbio.ac.cr/SSTN-IABIN/welcome.htm">http://ara.inbio.ac.cr/SSTN-IABIN/welcome.htm</a></td>
</tr>
<tr>
<td>Departamento de Zoologia, I.B., UNESP, Sao Paulo</td>
<td>Célio F. B. Haddad</td>
<td>NA</td>
</tr>
<tr>
<td>Field Museum of Natural History, Chicago</td>
<td>Harold K. Voris</td>
<td><a href="http://www.fieldmuseum.org/research_collections/zoology/">http://www.fieldmuseum.org/research_collections/zoology/</a></td>
</tr>
<tr>
<td>George Washington University</td>
<td>Guillermo Orti</td>
<td><a href="http://www.gwu.edu/~biology/faculty/ori.cfm">http://www.gwu.edu/~biology/faculty/ori.cfm</a></td>
</tr>
<tr>
<td>Inst of Chemical Biology and Fundamental Medicine, SB RAS</td>
<td>Alexander S. Grachodatsky</td>
<td><a href="http://www.niboch.nsc.ru/eng_index.html">http://www.niboch.nsc.ru/eng_index.html</a></td>
</tr>
<tr>
<td>Institut de Recherche pour le Développement, Paris, France</td>
<td>Philippe Gaubert</td>
<td><a href="http://pgaubert.perso.neuf.fr/">http://pgaubert.perso.neuf.fr/</a></td>
</tr>
<tr>
<td>Institute of Molecular and Cell Biology, Singapore</td>
<td>Byrappa Venkatesh</td>
<td><a href="http://www.imcb.a-star.edu.sg/php/venkatesh.php">http://www.imcb.a-star.edu.sg/php/venkatesh.php</a></td>
</tr>
<tr>
<td>Instituto Nacional de Cancer, Genetics Division</td>
<td>Hector N Seuanez</td>
<td><a href="http://www.inca.gov.br/">http://www.inca.gov.br/</a></td>
</tr>
<tr>
<td>Kunming Institute of Zoology, Chinese Academy of Sciences</td>
<td>Ya-ping Zhang</td>
<td><a href="http://www.kiz.ac.cn/en/">http://www.kiz.ac.cn/en/</a></td>
</tr>
<tr>
<td>LIRANS Institute, University of Bedfordshire, UK</td>
<td>David Michael Rawson</td>
<td><a href="http://www.beds.ac.uk/research/lirans/personnel/rawson_d">http://www.beds.ac.uk/research/lirans/personnel/rawson_d</a></td>
</tr>
<tr>
<td>LSU Museum of Natural Science</td>
<td>Frederick H. Sheldon</td>
<td><a href="http://app003.lsu.edu/natsci/mns.sf/$Content/Sheldon/OpenDocument">http://app003.lsu.edu/natsci/mns.sf/$Content/Sheldon/OpenDocument</a></td>
</tr>
<tr>
<td>Marine Mammal Institute, Oregon State University</td>
<td>C. Scott Baker</td>
<td><a href="http://mni.oregonstate.edu/">http://mni.oregonstate.edu/</a></td>
</tr>
<tr>
<td>Monterey Bay Aquarium Research Institute</td>
<td>Robert C. Vrijenhoek</td>
<td><a href="http://www.mbari.org/molecular">http://www.mbari.org/molecular</a></td>
</tr>
<tr>
<td>Museo Nacional de Ciencias Naturales (MNCN), Madrid</td>
<td>David R. Vieites</td>
<td><a href="http://www.vieiteslab.com">http://www.vieiteslab.com</a></td>
</tr>
<tr>
<td>Museu de Zoologia da Universidade de São Paulo</td>
<td>Hussam Zaher</td>
<td><a href="http://www.mz.usp.br/">http://www.mz.usp.br/</a></td>
</tr>
<tr>
<td>Museum of Vertebrate Zoology, UC Berkeley</td>
<td>Jimmy A. McGuire</td>
<td><a href="http://mvz.berkeley.edu/">http://mvz.berkeley.edu/</a></td>
</tr>
<tr>
<td>Natural History Museum of Los Angeles County</td>
<td>To be determined</td>
<td><a href="http://www.nhm.org/research">http://www.nhm.org/research</a></td>
</tr>
<tr>
<td>Ocean Park Corporation, Hong Kong</td>
<td>Paolo Martelli</td>
<td><a href="http://www.oceanpark.com.hk/">http://www.oceanpark.com.hk/</a></td>
</tr>
<tr>
<td>Pontifícia Universidade Católica do Rio Grande do Sul</td>
<td>Sandro Bonatto</td>
<td>NA</td>
</tr>
<tr>
<td>Royal Ontario Museum (ROM)</td>
<td>Robert W. Murphy</td>
<td><a href="http://labs.ceb.utoronto.ca/murphy/">http://labs.ceb.utoronto.ca/murphy/</a></td>
</tr>
<tr>
<td>San Diego Zoo’s Institute for Conservation Research</td>
<td>Oliver A. Ryder</td>
<td><a href="http://www.sandiegozoo.org/conservation/about/administrators/oliver_ryder_ph.d/">http://www.sandiegozoo.org/conservation/about/administrators/oliver_ryder_ph.d/</a></td>
</tr>
</tbody>
</table>
1200 to 1300 genera distributed in 3 major lineages: monotremes (platypus and echidnas—5 species), marsupials (~330 species, including the koala, kangaroos, and opossums), and the species-rich eutherian or placental mammals (~5000 species) (Nowak 1999; Wilson and Reeder 2005), (Table 1, Figure 2).

The G10K collection contains exemplars of 145 out of the 150 families (Supplementary Material, Appendix 2, mammals). At present, we have access to ~90% of nonmuroid and nonsciurid rodent genera and nonvespertilionid bat genera. Ultimately, we will target all 1200 to 1300 genera. Additional sampling will be applied to deeply divergent, and especially endangered, or Evolutionary Distinct and Endangered species (ZSL 2009), currently including all species of Zaglossus (echidna), Cuban and Hispaniolan Solenodon, Malayan Tapir (Tapirus indicus), aardvark (Orycteropus), and others. For fundamental biological investigation, another high priority is to sequence species exhibiting extreme phenotypes, such as deep-sea divers, long-lived species, high-altitude species, and species with distinct sensory modalities, such as echolocation. Our ultimate goal is to include within the collection species spanning the range of brain size, body size, and morphological convergence: aquatic species, gliders, lifespan extremes, nocturnals/diurnals, and social versus solitary species with diverse mating systems and varying levels of paternal care. We will also sample domestic animal species that have undergone recent and rapid evolution and contrast them to their counterpart wild species.

Capturing wide ecological diversity holds great potential for identifying the genomic changes underlying the major mammalian anatomical and behavioral transformations, including the evolution of advanced social and eusocial systems. Determining the genomic infrastructure for extreme physiological responses provides a unique opportunity for understanding the limits of mammalian tissues from resistance to disease to the ability to adapt to environmental disturbance.
Like eutherian mammals, living birds arose in the mid-Cretaceous (~100 Ma). Since then, birds have dispersed across the globe and now occupy most of Earth’s habitats and ecosystems representing a wide array of lifestyles. At this time, we know very little about the genetic and developmental underpinnings of this biological diversity, as high-quality genome sequences are available for only 2 species, the chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*). We expect that many key questions can and will be addressed as additional whole-genome sequences are accumulated and interpreted in the context of an increasingly accurate comparative framework (Hackett et al. 2008).

During recent decades, the avian systematics community has built large collections that house high-quality genetic samples of a substantial portion of avian diversity. These collections provide an essential resource for future genomic analyses of avian structural, functional, and behavioral diversity. With representation from 15 natural history collections distributed globally, the G10K collection includes specimens from 94% of the 34 orders, 91% of the 199 families, 73% of the 2172 genera, and 52% of the 9723 species of birds (Table 1, Figure 3). Every order is represented in multiple biospecimen collections, as are all but 17 families and all but 585 genera, ensuring at least 1 sample of high quality. We plan to sequence both sexes for a number of lineages, including the ratite birds, which like many avian species are externally monomorphic and, additionally, have relatively undifferentiated sex chromosomes.

Sampling each genus may result in oversampling of some avian orders and families (such as the extremely diverse passerines and hummingbirds), but we will strive to capture maximal phylogenetic coverage across the avian tree.
Figure 2. Consensus phylogeny of the major lineages of mammals. Topology and dates (Ma) are consensus estimates derived from Hedges and Kumar (2009) and included citations. Following the common names of taxon groups in parentheses is the number of living species for that group, followed by the number of G10K species with specific biospecimens nominated for G10K whole-genome sequence.
Figure 3. Consensus phylogeny of the major lineages of birds. Topology and dates (Ma) are derived from combined-data tree reported in Hackett et al. (2008), rendered ultrametric by nonparametric rate-smoothing (Sanderson 1997) and scaled to a root age of 119 Ma based on an average of multiple dating studies (van Tuinen et al. 2006). Following the common names of taxon groups in parentheses is the number of living species for that group followed by the number of G10K species with specific biospecimens nominated for G10K whole-genome sequence.
Nonavian reptile diversity includes snakes, lizards, turtles, crocodilians, and 2 species of tuatara. Because the traditional view of interfamilial relationships (based on morphology) differs appreciably from recent molecular phylogenies and the molecular phylogenies differ from one another, major issues such as the origin of snakes (which are clearly nested within lizards) remain controversial (Fry et al. 2006; Vidal and Hedges 2009). In addition to these uncertainties, the phylogenetic relationships within and among the major groups of reptiles (i.e., families) are often uncertain, for example, among the “colubroid” snakes (Hedges et al. 2009; Zaher et al. 2009) and species-rich assemblages of lizards. Major revisions have occurred within many groups, such as the geckos, where additional families are now recognized (Gamble et al. 2008). Following online databases including the TIGR Reptile Database (Uetz 2009), reptile diversity is distributed among the following groupings: Snakes are divided among 18 families, 484 genera, and 3313 species; lizards comprise 30 families, 499 genera, and 5351 species; and turtle diversity is divided among 13 families, 94 genera, and 313 species (Turtle Taxonomy Working Group 2007). Crocodiles include 23 species divided among 9 genera in 3 taxonomic families. And the 2 species of tuatara are the only extant members of the formerly diverse and widespread Rhynchocephalia. Total reptile diversity therefore includes 65 families, 1087 genera, and 9002 species. The G10K collection has 97%, 69%, and 37% of these, respectively (Table 1, Figure 4). In addition to these DNA and tissue samples, substantial BAC-library resources are available for nonavian reptiles that could facilitate the G10K project (Wang et al. 2006).
Amphibians

The Class Amphibia is divided into 3 orders: Anura (frogs), Caudata (salamanders), and Gymnophiona (caecilians), derived from a common ancestor 300 Ma and representing the only 3 surviving lineages from a much greater diversity that existed before the Permian extinction 250 Ma (Marjanovic and Laurin 2007). These major clades contain 5811 frog species, 583 salamander species, and 176 caecilian species, respectively (AmphibiaWeb 2009). Amphibian taxonomy is currently in a state of flux, with many new proposed taxonomic changes resulting from molecular phylogenetic analyses. Although controversial, we summarize amphibian diversity and tissue holdings for higher taxonomic groups (Supplemental Material, Appendix 2, amphibians) following the AmphibiaWeb (2009) database. This taxonomy contains 56 families of amphibians shared among the 3 orders, containing a total of 510 genera and 6570 species. The G10K collection contains a total of 1760 species (27%), 301 genera (59%), and 50 families (89%) (Table 1, Figure 5).

Amphibians are notorious for their morphological homoplasy due to developmental constraints (Shubin et al. 1995) as well as spectacular adaptive convergences in morphology (Bossuyt and Milinkovitch 2000), behavior, and development, for example, roughly 15 independent evolutionary origins of direct development from an ancestral biphasic life history (Hanken et al. 1997). Perhaps the most striking example is the convergent evolution in toxicity, coloration, and parental care between mantellid frogs of Madagascar and dendrobatid frogs in the Neotropics, as well as repeated parallel evolution of these traits within each of these 2 taxonomic families (Vences et al. 2003; Chiari et al. 2004). Such homoplasies have wreaked havoc on amphibian taxonomy, but offer marvelous opportunities to study the
genetic basis of the repeated evolution of complex traits involved in both morphological and behavioral evolution.

Collectively, amphibians are of global conservation concern, most recently because of a rapid decline in populations and disappearance of species (Mendelson et al. 2006). A chytrid fungus, \textit{Batrachochytrium dendrobatidis}, has been implicated in these declines (James et al. 2009), but habitat loss, pollutants, pesticides, herbicides, fertilizers, and climatic changes are also factors of concern. In the face of such diversity crises, sequencing many species of amphibians has enormous potential to provide insight into novel antimicrobial compounds, given that many species of frogs harbor a diverse array of such compounds (Zasloff 2002; Vanhoey et al. 2003). The same antimicrobial peptide sequence is rarely recovered from closely related species. Genomic approaches to searching for such antimicrobial diversity using stem cell lines, transcriptomes, and whole-genome sequencing are clearly warranted.

**Fishes**

Fishes include all non-tetrapod vertebrates comprising 1) jawless vertebrates (hagfishes and lampreys, 114 species), 2) chondrichthyans (sharks, rays, and chimaeras, \~1200 species), 3) actinopterygians (ray-fin fishes, \~30 000 species), and 4) piscine sarcopterygians (coelacanths, lungfishes, 8 species). Total described diversity comprises approximately 31 500 species (Eschmeyer 1998), but actual diversity is probably greater than 50 000 species. A broad outline of the evolution of these most deeply branching of the vertebrate clades is provided by Stiassny et al. (2004).

Fishes account for nearly 50% of all described living species of vertebrates, exhibiting a vast diversity in their morphology, physiology, behavior, and ecological adaptations and providing an exceptional opportunity to study basic vertebrate biology. Fishes are also important as a food source for human consumption totaling about US $51 billion in trade in 2001 (Tidwell and Allan 2001). In 2006, global capture fisheries were estimated at US $91 billion and global aquaculture (including invertebrates) at US $79 billion (FAO 2008). There is also huge global recreational spending. Fishery activities of all types probably total in excess of US $200 billion per year (FAO 2008). Some 16% of all human protein consumption is fish protein, and about 1 billion people depend on fishes as their major source of protein. Because of the great demand, many groups of fishes are overexploited. Molecular data for commercially important species of fishes, especially those that are currently endangered and those raised by aquaculture, will be valuable in designing strategies for maintaining sustainable stocks and combating disease and other threats.

Fish tissues for the G10K project reside in a number of institutions and are usually curated as parts of formal institutional collections. The total number of species represented by tissue samples is not known precisely, but 6,400 species have been DNA barcoded and collections of new species continue to be added (Wiley E, KU, personal communication). Fresh material from many commonly available species can be obtained easily from fishing boats and the pet-trade industry for both genome and other molecular projects. The G10K project has in hand suitable samples from 62/62 orders (100%), 424/532 families (80%), 1,777 of about 4,956 genera (36%), and 4,246 of about 31,564 named species (13%) (Table 1, Figure 1). We have identified other partner institutions that are anticipated to provide a minimum of 2,500 additional species that will be officially incorporated into the project.

The largest known animal genome is that of the marbled lungfish, \textit{Protopterus aethiopicus}, with a haploid size of 133 pg (about 130 Gbp), followed by the salamanders \textit{Necturus lewisi} and \textit{Necturus punctatus} at 120 pg (about 117 Gbp) (Gregory 2009). The genomes are bloated through the activity of transposons that, combined with their enormous size, make genome sequencing and assembly extremely challenging. Although RNA sequencing is one avenue by which we may get direct access to interesting biology in these species, we nevertheless recommend that full-genome sequencing projects be undertaken for large-genome species. There are important questions pertaining to gene regulation, genome structure, and genome evolution that cannot be answered from analysis of transcribed RNA alone.

**Discussion**

Careful observations of the morphological and functional adaptations in vertebrates have formed the basis of biological studies for a millennium, but it is only recently that we have been able to observe the action of evolution directly at the genetic level. It is not known whether convergent adaptations in independent lineages are often governed by analogous changes in a small number of orthologous genome loci or if macroevolutionary events in separate lineages usually result from entirely idiosyncratic combinations of mutations. The evidence from several recent studies points toward the former hypothesis (Eizirik et al. 2003; Nachman et al. 2003). For example, adaptive hind-limb reduction occurred independently many times in different lineages and even within the same species, just as sticklebacks in different lakes adapted from an oceanic to a freshwater environment (Shapiro et al. 2006). These stickleback adaptations are all traced to independent deletions of the same distal enhancer of the \textit{PITX2} development gene, demonstrating remarkable convergent evolution at the genomic level (Kingsley D, HHMI, personal communication). By cataloging the footprints of adaptive evolution in every genomic locus on every vertebrate lineage, the G10K project will provide the power to thoroughly test the “same adaptation, same loci” hypothesis, along with other fundamental questions about molecular adaptive mechanisms.

In the course of this investigation, we will discover the genetic loci governing fundamental vertebrate processes. The study of the evolution of viviparity is an outstanding example. Birds, crocodiles, and turtles all lay eggs, whereas
In all these ways, the G10K project will engage the public in the quest for the scientific basis of animal diversity and in the application of the knowledge we gain to halt extinctions and improve animal health.

As the printing of the first book by Johannes Gutenberg altered the course of human history, so did the human genome project forever change the course of the life sciences with the publication of the first full vertebrate genome sequence. When Gutenberg’s success was followed by the publication of other books, libraries naturally emerged to hold the fruits of this new technology for the benefit of all who sought to imbibe the vast knowledge made available by the new print medium. We must now follow the human genome project with a library of vertebrate genome sequences, a genomic ark for thriving and threatened species alike, and a permanent digital record of countless molecular triumphs and stumbles across some 600 million years of evolutionary episodes that forged the “endless forms most beautiful” that make up our living world.

**Supplementary Material**


**Appendix**

**Genome 10K Community of Scientists (G10KCOS)**

**Authors**

Coordinators and corresponding authors: David Haussler (Howard Hughes Medical Institute, UCSC, CBSE/TITI E2501, University of California, Santa Cruz, Santa Cruz, CA, e-mail: haussler@soe.ucsc.edu); Stephen J. O’Brien (National Cancer Institute, Laboratory of Genomic Diversity, Frederick, MD, e-mail: stephen.obrien@nih.gov); Oliver A. Ryder (San Diego Zoo’s Institute for Conservation Research, Escondido, CA, e-mail: oryder@sandiegozoo.org)

Coauthors: committee chairs (alphabetical): F. Keith Barker (University of Minnesota, Department of Ecology, Evolution and Behavior, St Paul, MN); Michele Clamp (The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA); Andrew J. Crawford (Universidad de los Andes, Departamento de Ciencias Biológicas, Carrera 1E No. 18A–10, A.A. 4976, Bogotá, Colombia); Robert Hanner (Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada); Olivier Hanotte (The Frozen Ark Project—University of Nottingham, School of Biology, University Park, Nottingham, Nottinghamshire, UK); Warren E. Johnson (National Cancer Institute, Laboratory of Genomic Diversity, Frederick, MD); Jimmy A. McGuire (Museum of Vertebrate Zoology and Department of Integrative Biology, University of California, Berkeley, CA); Webb Miller (The Pennsylvania State University, Biology Department, 208 Mueller Laboratory, University Park, PA); Robert W. Murphy (Royal Ontario Museum, Department of Natural History, 100 Queen’s Park, Toronto, ON, Canada); William J. Murphy (Texas A&M University,
and Evolutionary Biology, Box 951606, 2312 LSB, Los Angeles, CA; Terrie M. Williams (University of California Santa Cruz, Center for Ocean Health—Department of Ecology and Evolutionary Biology, Santa Cruz, CA); Nathan D. Wolfe (Global Viral Forecasting Initiative, One Market, Spear Tower, Suite 3574, San Francisco, CA and Stanford University, Program in Human Biology, Stanford, CA); Ya-Ping Zhang (State Key Laboratory of Genomic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 32 Jiaoachagdong ST, Kunming, Yunnan, China)

Committees:

Birds Group Members: *F. Keith Barker, Joel Cracraft, Scott V. Edwards, Olivier Hanotte, *Frederick H. Sheldon


*Chairs

Funding
American Genetic Association, Gordon and Betty Moore Foundation, NHGRI Intramural Sequencing Center, and UCSC Alumni Association to cost of the Genome 10K workshop; Howard Hughes Medical Institute to D.H.; Gordon and Betty Moore Foundation to S.C.S.; Assembling the Euteleost Tree of Life to E.W.; National Science Foundation (0732819 to E.W., DEB-0640967 and 0543556 to J.A.M., 0817042 to H.B.S., EF0629849 to W.J.M., DEB-0443470 to G.O.); The Global Viral Forecasting Initiative to N.W., B.P., and M.L.; Biomedical Research Council of A*STAR, Singapore to B.V.; Natural Sciences and Engineering Research Council Discovery Grant to R.W.M.; National Basic Research Program of China (973 Program, 2007CB411600), the National Natural Science Foundation of China (30621092), and Bureau of Science and Technology of Yunnan Province to Y.Z.; MCB and SB RAS Programs (A.S.G.); Portuguese-American Foundation for Development, CIBIO, UP, University of Montana [G.L.] and Portuguese Science Foundation [PTDC/CVT/69438/2006; PTDC/BLA-BDE/65625/2006 to G.L.]

Acknowledgments
We wish to thank R. Fuller and S. Karl for project assistance and our reviewers for helpful comments.

References


