Journal of Heredity 2009:100(6):659–674
doi:10.1093/jhered/esp086
Advance Access publication November 5, 2009

© The American Genetic Association. 2009. All rights reserved.
For permissions, please email: journals.permissions@oxfordjournals.org.

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 10^8 for all metazoans and approximately 6×10^4 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 10^8 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 (Hoekstra and Coyne 2007; Stranger et al. 2007). Almost no part of the known noncoding vertebrate gene regulatory apparatus bears any discernable resemblance at the DNA level to analogous systems in our deuterostome distant cousins. Yet, noncoding DNA segments represents 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

	Orders			Families			Genera			Species		
Groups	With G10K samples	Total	% of total	With G10K samples	Total	% of total	With G10K samples	Total	% of total	With G10K samples	Total	% of total
Mammals	27	27	100	145	150	97	763	1230	62	1826	5416	34
Birds	32	34	94	182	199	91	1587	2172	73	5074	9723	52
Amphibians	3	3	100	50	56	89	301	510	59	1760	6570	27
Reptiles	4	4	100	63	65	97	751	1087	69	3297	9002	37
Fishes	62	62	100	424	532	80	1777	4956	36	4246	31 564	13
Totals	128	130	98	864	1002	86	5179	9955	52	16 203	62 275	26

Species and other taxa numbers are initially from NCBI taxonomy (www.ncbi.nlm.nih.gov) (Wheeler et al. 2008), as specified by Wilson and Reeder (2005) for the mammals, Hackett et al. (2008); Howard et al. (2003) for birds, Marjanovic and Laurin (2007); AmphibiaWeb (2009) for Amphibia, Catalog of Fishes (Eschmeyer 1998) for fishes, The TIGR Reptile Database (Uetz et al. 2007) for reptiles.

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

Institutions	Steward(s)	Web address
Academy of Natural Sciences, Philadelphia	Nate Rice	http://www.ansp.org/
American Museum of Natural History	Joel Cracraft	http://research.amnh.org/ornithology/personnel/jlc.htm
Australian National University	Jennifer A. Marshall Graves	http://www.rsbs.anu.edu.au/ResearchGroups/CGG/index.php
Australian National Wildlife Collection, Canberra	Leo Joseph	http://www.csiro.au/places/ANWC.html
Bell Museum of Natural History, University of Minnesota	F. Keith Barker	http://www.bellmuseum.org/
Biodiversity Institute of Ontario	Alexei Borisenko	http://www.biodiversity.ca
Biodiversity Research Institute, University of Kansas	Edward O. Wiley	http://www.nhm.ku.edu/fishes/
Burke Museum, University of Washington	To be determined	http://www.washington.edu/burkemuseum/collections/genetic/index.php
California Academy of Sciences	Jens V. Vindum	NA
CIBIO, University of Porto, Portugal	Albano Beja-Pereira	http://cibio.up.pt/
Círculo Herpetológico de Panamá	Roberto Ibáñez	http://ara.inbio.ac.cr/SSTN-IABIN/welcome.htm
CSIRO Marine and Atmospheric Research	Alastair Graham	http://www.cmar.csiro.au/anfc/
Departamento de Zoologia, I.B., UNESP, Sao Paulo	Célio F. B. Haddad	NA
Field Museum of Natural History, Chicago	Harold K. Voris	http://www.fieldmuseum.org/research_collections/zoology/
George Washington University	Guillermo Orti	http://www.gwu.edu/~biology/faculty/orti.cfm
Inst of Chemical Biology and Fundamental Medicine, SB RAS	Alexander S. Graphodatsky	http://www.niboch.nsc.ru/eng_index.html
Institut de Recherche pour le Développement, Paris, France	Philippe Gaubert	http://pgaubert.perso.neuf.fr/
Institute of Molecular and Cell Biology, Singapore	Byrappa Venkatesh	http://www.imcb.a-star.edu.sg/php/venkatesh.php
Instituto Nacional de Cancer, Genetics Division	Hector N Seuanez	http://www.inca.gov.br/
Kunming Institute of Zoology, Chinese Academy of Sciences	Ya-ping Zhang	http://www.kiz.ac.cn/en/
LIRANS Institute, University of Bedfordshire, UK,	David Michael Rawson	http://www.beds.ac.uk/research/lirans/personnel/rawson_d
LSU Museum of Natural Science	Frederick H. Sheldon	http://appl003.lsu.edu/natsci/lmns.nsf/$Content/Sheldon?OpenDocument
Marine Mammal Institute, Oregon State University	C. Scott Baker	http://mmi.oregonstate.edu/
Monterey Bay Aquarium Research Institute	Robert C. Vrijenhoek	http://www.mbari.org/molecular
Museo Nacional de Ciencias Naturales (MNCN), Madrid	David R. Vieites	http://www.vieiteslab.com
Museu de Zoologia da Universidade de São Paulo	Hussam Zaher	http://www.mz.usp.br/
Museum of Comparative Zoology, Harvard	Scott Edwards	http://www.mcz.harvard.edu
Museum of Vertebrate Zoology, UC Berkeley	Jimmy A. McGuire	http://mvz.berkeley.edu/
Museum Victoria, Australia	Joanna Sumner	http://museumvictoria.com.au/collections-research/our-research/sciences/staff/joanna-sumner/
National Cancer Institute Lab of Genomic Diversity	Stephen J. O'Brien	http://home.ncifcrf.gov/ccr/lgd/
Natural History Museum of Los Angeles County	To be determined	http://www.nhm.org/research
Ocean Park Corporation, Hong Kong	Paolo Martelli	http://www.oceanpark.com.hk/
Pontificia Universidade Catolica do Rio Grande do Sul	Sandro Bonatto	NA
Royal Ontario Museum (ROM)	Robert W. Murphy	http://labs.eeb.utoronto.ca/murphy/
San Diego Zoo's Institute for Conservation Research	Oliver A. Ryder	http://www.sandiegozoo.org/conservation/about/administrators/oliver_ryder_ph.d/

Table 2. Continued

Institutions	Steward(s)	Web address
Smithsonian Institution, National Museum of Natural History	Roy W. McDiarmid	http://vertebrates.si.edu/herps/
Smithsonian Tropical Research Institute	Eldredge Bermingham	http://www.stri.org/
South Australian Museum	Steve Donnellan	http://www.samuseum.sa.gov.au/page/default.asp?site=1&id=1307
Southwest Fisheries Science Center, NMFS	Gabriela Serra-Valente	http://swfsc.noaa.gov/textblock.aspx?Division%20=%20PRD%20=%20229%20=%2012498
Swedish Museum of Natural History	Per Ericson	http://www.nrm.se/en/menu/researchandcollections/departments/vertebratezoology.74_en.html
Texas A&M University	William J. Murphy	http://gene.tamu.edu/faculty_pages/faculty_MurphyW.php
The Frozen Ark	Olivier Hanotte	http://www.frozenark.org/
The Global Viral Forecasting Initiative	Matthew LeBreton	www.gvfi.org
Universidade Federal do Rio de Janeiro, Genetics Dept.	Miquel Moreira & Cibele Bonvicino	http://www.inca.gov.br/conteudo_view.asp?id=414
University of Auckland, New Zealand, School of Biological Sciences	Rochelle Constantine & C. Scott Baker	http://www.sbs.auckland.ac.nz/
University College-Dublin	To be determined	http://www.ucd.ie/research/people/biologyenvscience/dremmacteeling/
University of California, Riverside	To be determined	http://www.biology.ucr.edu/people/faculty/Springer.html
University of California, Santa Cruz, Department of Ecology & Evolutionary Biology	Barry Sinervo	http://bio.research.ucsc.edu/~barrylab/
University of California, Santa Cruz, Mammal Physiology Program	Terrie M. Williams	http://bio.research.ucsc.edu/people/williams/
University of Kansas, Department of Ecology and Evolutionary Biology	Rafe Brown	http://www.nhm.ku.edu/rbrown/
University of Minnesota, Cell Biology & Development	Tony Gamble	http://www.tc.umn.edu/~gambl007/
University of Montana	Gordon Luikart	http://dbs.umt.edu/research_labs/allendorflab/
University of Sheffield	Terry Burke	http://www.shef.ac.uk/molecol/terry-burke
University of Texas at Arlington	Jonathan A. Campbell	http://biology.uta.edu/herpetology
Villanova University	Aaron Bauer	http://www.villanova.edu/artsci/biology/
Zoological Institute, Technical University of Braunschweig	Miguel Vences	http://www.mvences.de/
Zoological Museum of Copenhagen, Denmark	Jon Fjeldsa	http://zoologi.snm.ku.dk/english/

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.

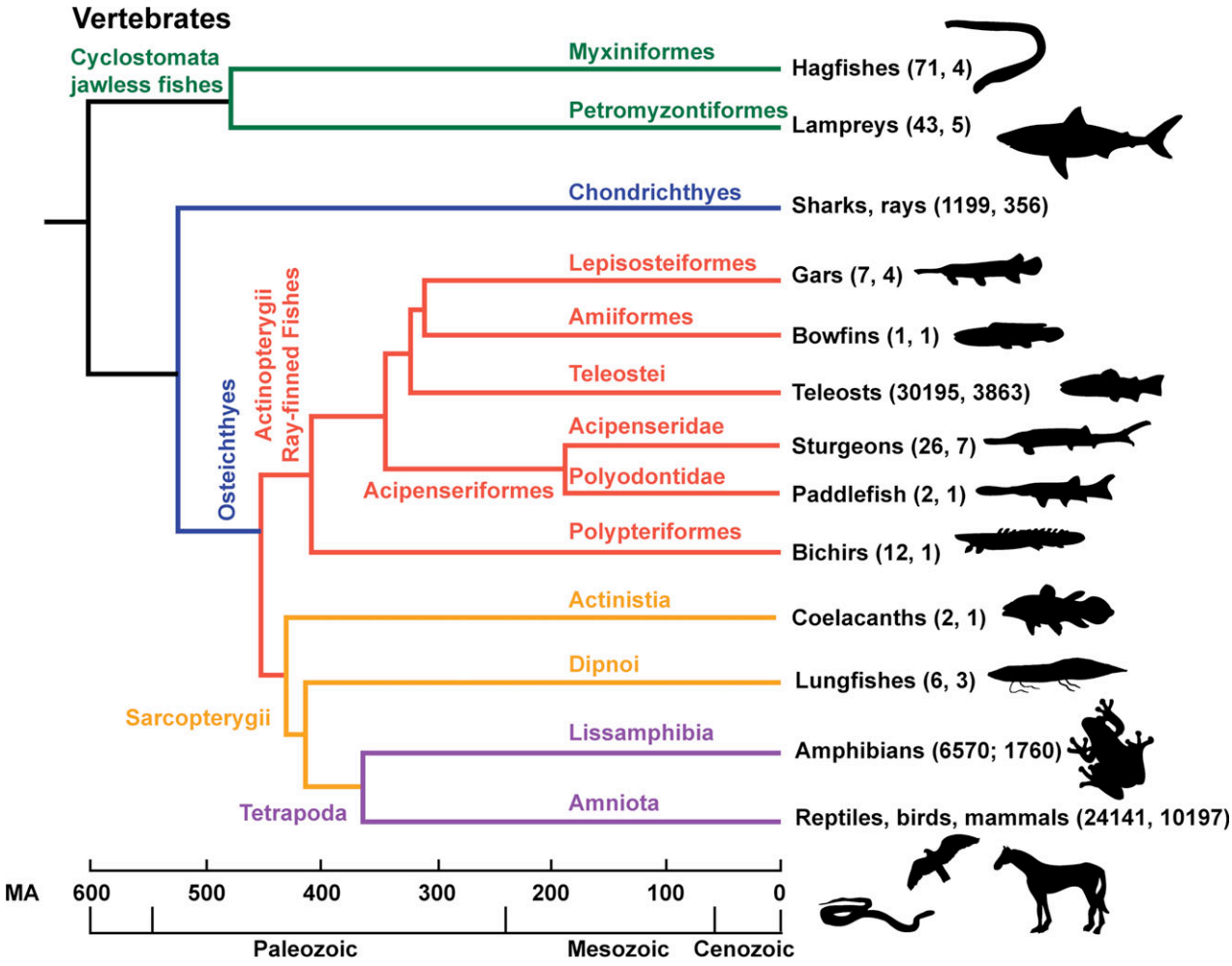

Figure 1. Consensus phylogeny of the major lineages of vertebrates. Topology and divergence dates (Ma) are consensus estimates derived from Hedges and Kumar (2009) and included citations and amended per Benton and Donoghue (2007), Janvier (2006), Maisey (2000), and Sansom et al. (1996). Following the 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.

Birds

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.

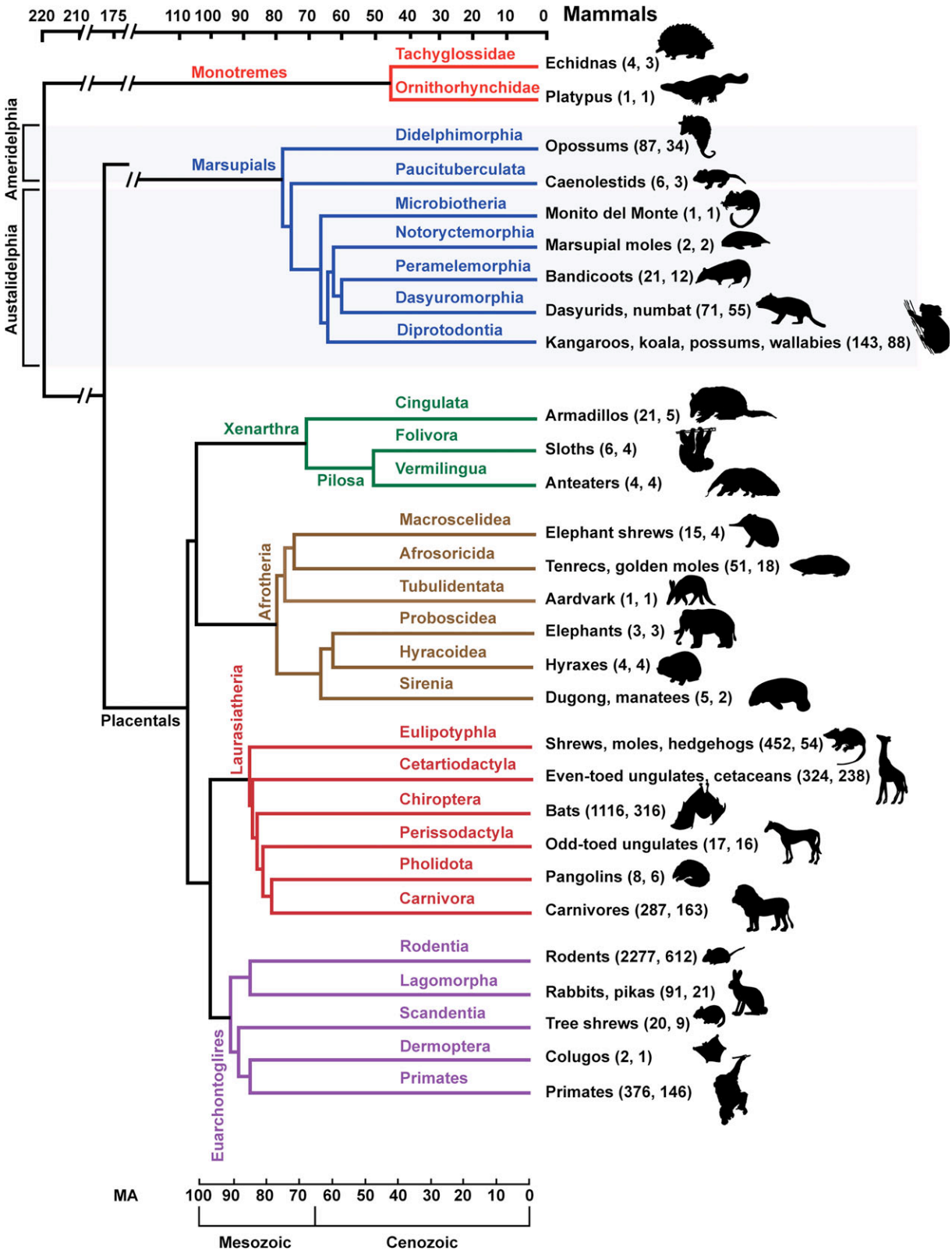

Figure 2. Consensus phylogeny of the major lineages of mammals. Topology and dates (Ma) are consensus estimates derived from Hedges and Kumar (2009) and included citations. Following the common names of taxon groups in parentheses is the number of living species for that group, followed by the number of G10K species with specific biospecimens nominated for G10K whole-genome sequence.

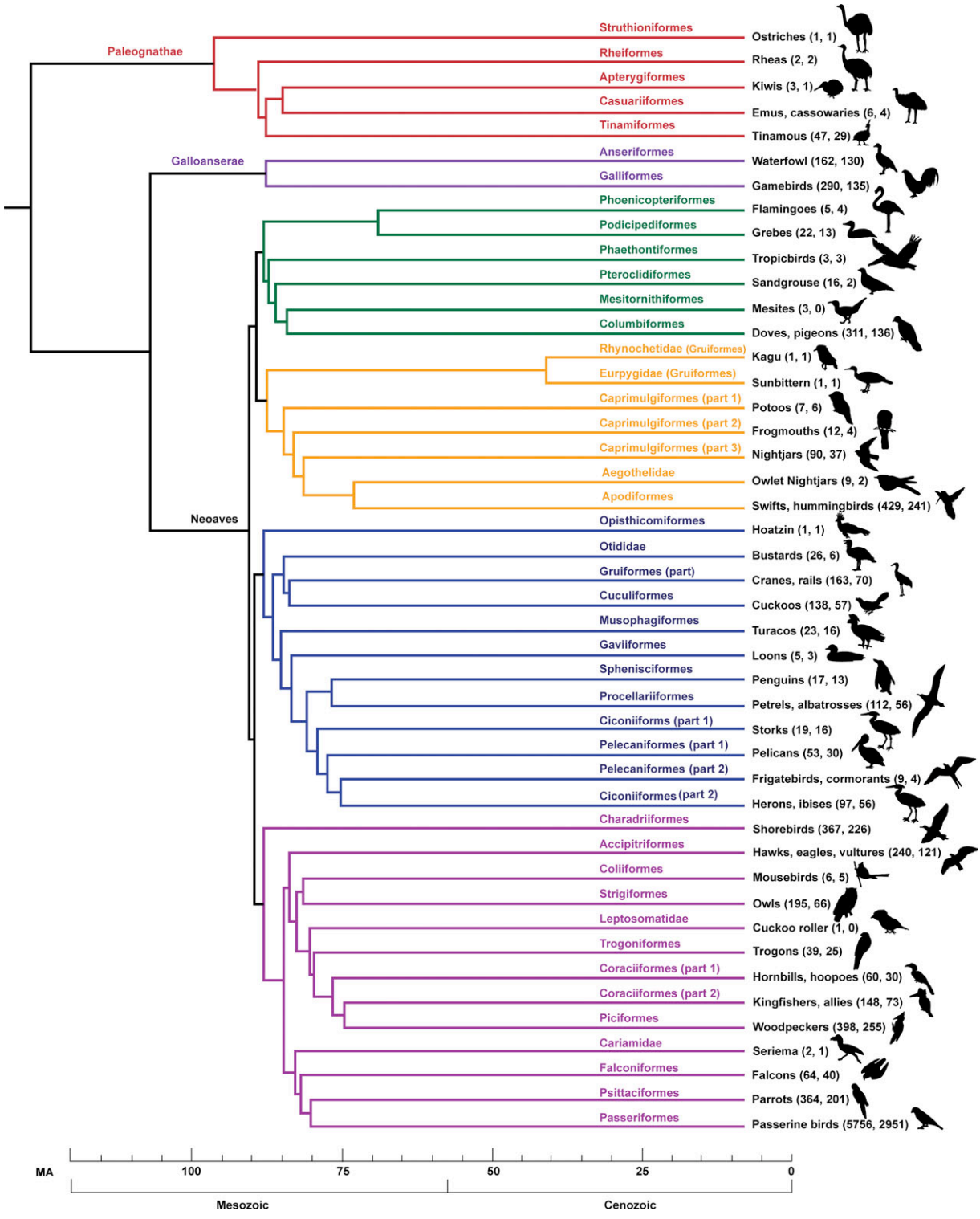

Figure 3. Consensus phylogeny of the major lineages of birds. Topology and dates (Ma) are derived from combined-data tree reported in Hackett et al. (2008), rendered ultrametric by nonparametric rate-smoothing (Sanderson 1997) and scaled to a root age of 119 Ma based on an average of multiple dating 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.

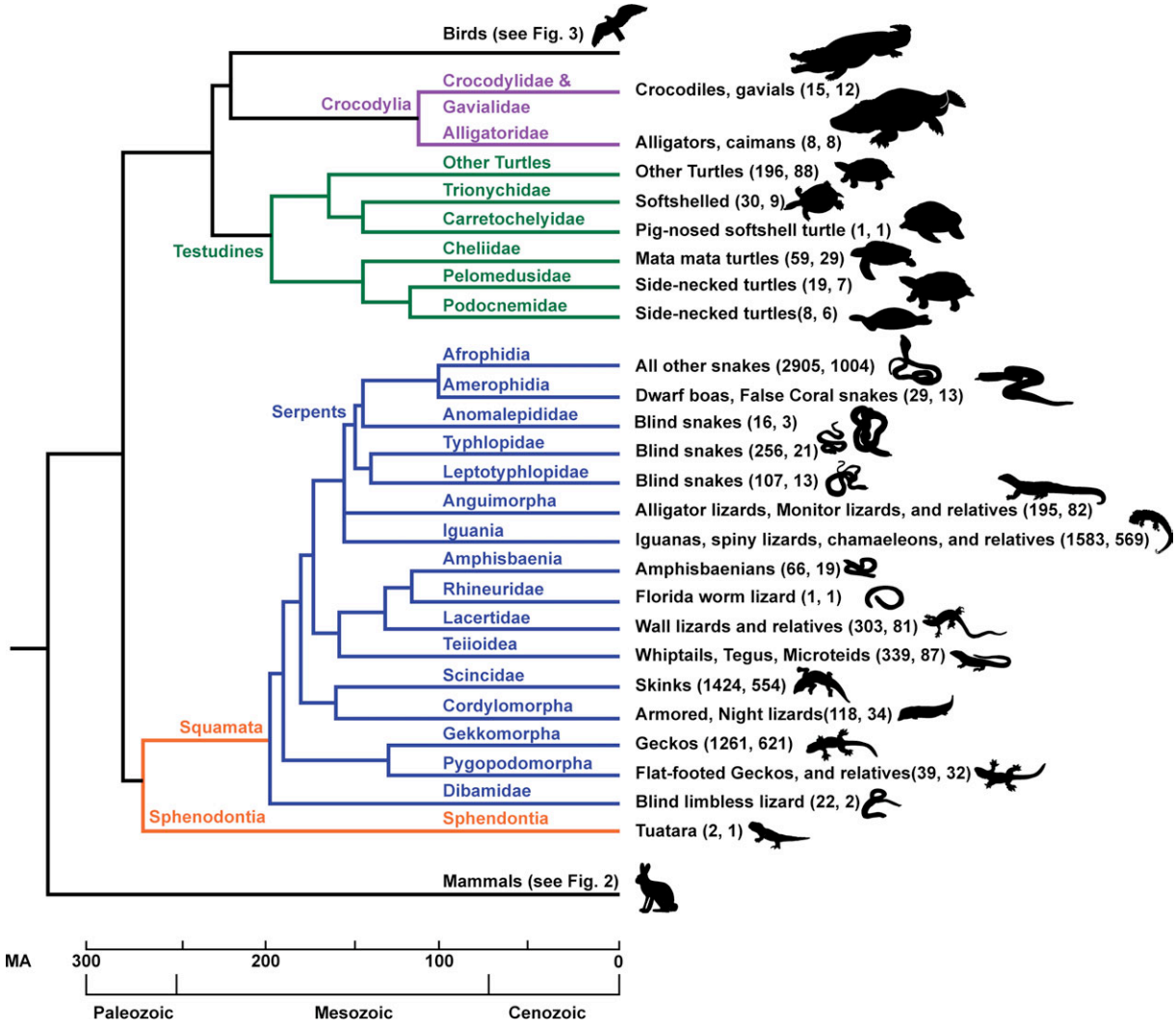

Figure 4. Consensus phylogeny of the major lineages of nonavian reptiles. Topology and dates (Ma) are consensus estimates derived from Hedges and Kumar (2009) and included citations. Following the common names of taxon groups in parentheses is the number of living species for that group followed by the number of G10K species with specific named biospecimens nominated for G10K whole-genome sequence.

Nonavian Reptiles

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 Rhyncocephalia. 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).

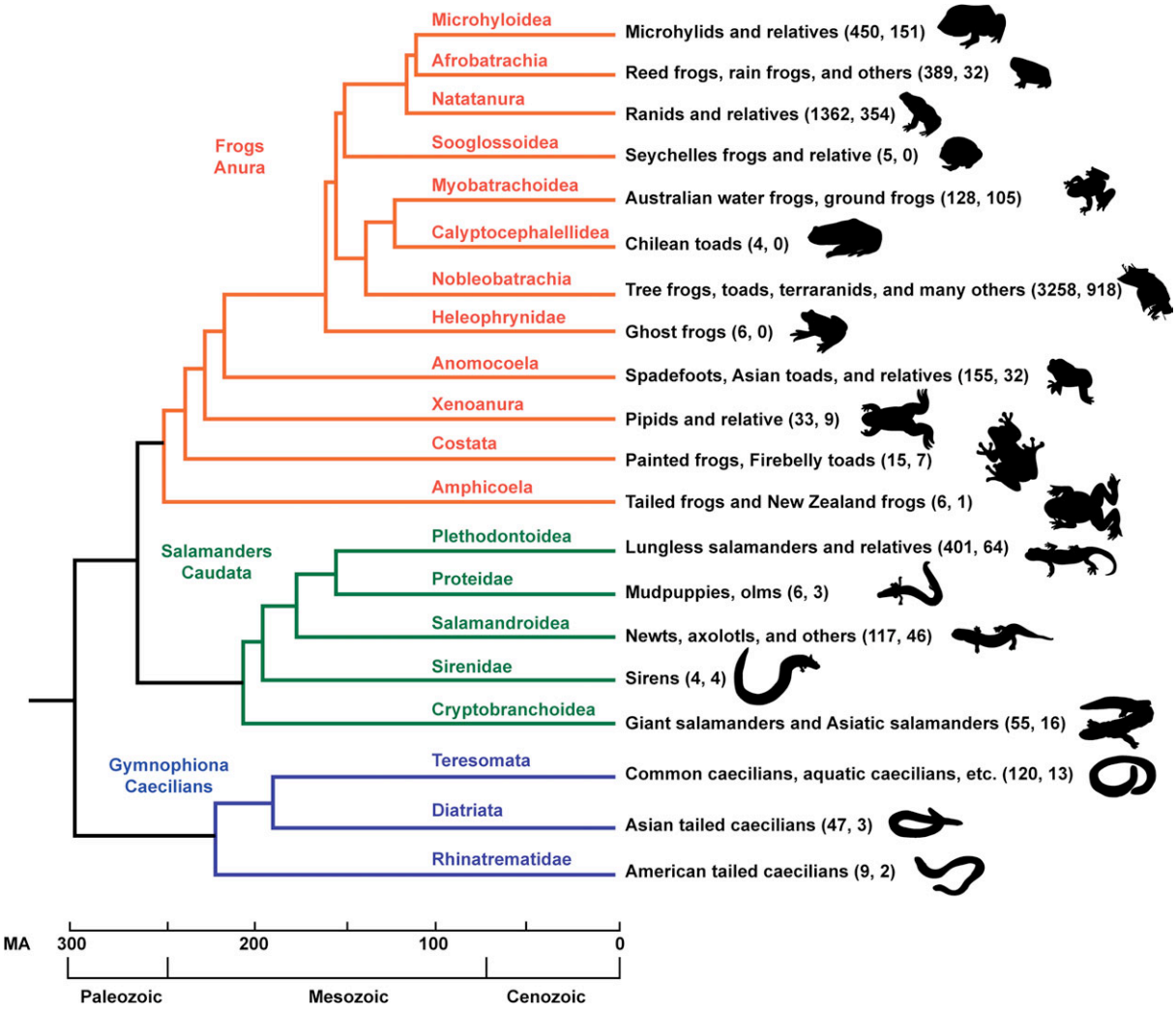

Figure 5. Consensus phylogeny of the major lineages of amphibians. Topology and dates (Ma) are consensus estimates derived from Hedges and Kumar (2009) and included citations. Following the common names of taxon groups in parentheses is the number of living species for that group followed by the list of G10K species with specific named biospecimens nominated for G10K whole-genome sequence.

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, *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; Vanhoye 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 nontetrapod 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%), 1777 of about 4956 genera (36%), and 4246 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 2500 additional species that will be officially incorporated into the project.

The largest known animal genome is that of the marbled lungfish, *Protopterus aethiopicus*, with a haploid size of 133 pg (about 130 Gbp), followed by the salamanders *Necturus lewisi* and *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 *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

apart from monotremes, mammals are all live bearers. Thus, there was one fundamental transition from oviparity to viviparity in these amniotes, which caused a fundamental reorganization in the developmental program and large-scale change in gene interactions that we are only just beginning to understand. Remarkably, however, nonavian reptiles have over 100 independent evolutionary origins of viviparity (Blackburn 2000). Fish have an equally spectacular variety of such transitions, along with some amphibians, such as the frog genus *Gastrotheca*, which includes species with placental-like structures (Duellman and Trueb 1986). These many independent instances of the evolution of viviparity afford an extraordinary opportunity to explore the genomics behind this reproductive strategy.

The architecture of sex determination in vertebrates is similarly diverse, with examples of XY, ZW, and temperature-dependent mechanisms. The G10K project thus provides an equally exciting opportunity for dissection of this diversity. In fact, a few vertebrate species have abandoned sex altogether. What happens when an asexual genome descends from an ancestral sexual genome, as has occurred repeatedly in *Aspidoscelis* lizard lineages? Are the independent parthenogenetic genomes parallel in any way? In one group of lizards, genus *Darevskia*, the formation of unisexual species is phylogenetically constrained (Murphy et al. 2000), yet in others, for example, *Aspidoscelis*, it is not. Many species of lizards and snakes are also known to have facultative parthenogenesis: Unmated females produce viable eggs and offspring. Unisexuality also occurs in amphibians and fishes by gynogenesis, hybridogenesis, and in amphibians by kleptogenesis (Bogart et al. 2007). Sequential hermaphrodite fishes can change their sex. Do these parallel convergent changes involve the same genes? The evolution of longevity remains another question of great interest. What mechanisms are responsible for the 2 orders of magnitude differences among vertebrates and what sets the limits for long-lived species found in each of the vertebrate clades? By identifying genomic loci that support different evolutionary innovations such as these, the data from the G10K project will drive fundamental progress in molecular and developmental biology.

The symphony of vertebrate species that cohabit on our planet attests to underlying life processes with remarkable potential. Genomics reveals a unity behind these life processes that is unrivaled by any other avenue of investigation, exposing the undeniable relatedness and common origin of all species. By revealing genetic vulnerabilities in endangered species and tracking host–pathogen coevolution, genomics also plays an increasing role in sustaining biodiversity and combating emerging infectious diseases. Thus, the information in the genomes of threatened and endangered species revealed by the G10K project will be crucial to conservation efforts (Ryder et al. 2000; O'Brien 2003; Ryder 2005; Kohn et al. 2006; Schwartz et al. 2009). In studying the genomes of recently extinct species as well, molecular aspects of species' vulnerability can be revealed and vital gaps in the vertebrate record restored. 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

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Appendix

Genome 10K Community of Scientists (G10KCOS) Authors

Coordinators and corresponding authors: David Haussler (Howard Hughes Medical Institute, UCSC, CBSE/ITI 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,

Department of Veterinary Integrative Biosciences, College Station, TX); Frederick H. Sheldon (Museum of Natural Science and Department of Biological Sciences, Louisiana State University, 119 Foster Hall, Baton Rouge, LA); Barry Sinervo (Department of Ecology and Evolutionary Biology University of California, Santa Cruz, Earth & Marine Sciences A308, Santa Cruz, CA); Byrappa Venkatesh (Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, Biopolis, Singapore, Republic of Singapore); Edward O. Wiley (Department of Ecology and Evolutionary Biology, University of Kansas, Natural History Museum and Biodiversity Research Center, Lawrence, KS)

Additional authors (alphabetical) Fred W. Allendorf (Division of Biological Sciences, University of Montana, Missoula, MT); George Amato (Center for Conservative Genetics, American Museum of Natural History, New York, NY); C. Scott Baker (Marine Mammal Institute and Department of Fisheries and Wildlife, Oregon State University, Newport, OR); Aaron Bauer (Department of Biology, Villanova University, Mendel Hall Rm 191, Biology, Villanova, PA); Albano Beja-Pereira (CIBIO, Campus Agrario de Vairao, R., Maonte-Crasto, Vairao, Portugal); Eldredge Bermingham (Smithsonian Tropical Research Institute, PO Box 0843-03092, Balboa, Ancon Panama-Republic Of Panama); Giacomo Bernardi (University of California Santa Cruz, Department of Ecology and Evolutionary Biology, Santa Cruz, CA); Cibele R. Bonvicino (Genetics Division, Instituto Nacional de Câncer, Rua André Cavalcanti, 37, 4o andar, Rio de Janeiro, RJ 20231-050, and Laboratório de Biologia e Parasitologia de Mamíferos Reservatórios Silvestres, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil); Sydney Brenner (Salk Institute for Biological Studies, PO Box 85800, San Diego, CA, Okinawa Institute of Science and Technology, Okinawa, Japan); Terry Burke (Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK); Joel Cracraft (American Museum of Natural History, Department of Ornithology, New York, NY); Mark Diekhans (University of California, Santa Cruz, Santa Cruz, CA); Scott Edwards (Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA); Per G.P. Ericson (Swedish Museum of Natural History, PO Box 50007, Stockholm, Sweden); James Estes (Department of Ecology and Evolution, Center for Ocean Health, Santa Cruz, CA); Jon Fjelsda (University of Copenhagen, Zoologisk Museum, Universitetsparken 15, Museet—Bygn.11, 2-4-465, Denmark); Nate Flesness (ISIS, International Species Information System, Minneapolis, MN); Tony Gamble (University of Minnesota, Department of Genetics, Cell Biology, Room 6-160 Jackson Hall, Minneapolis, MN); Philippe Gaubert (Muséum National d'Histoire Naturelle, UMR BOREA IRD 207, 43 rue Cuvier—CP 26, Paris, France); Alexander S. Graphodatsky (Institute of Chemical Biology and Fundamental Medicine, Russian Academy of Science, Siberian Branch, Prospect Lavrentieva, 10, Novosibirsk, Novosibirsk Region, Russia); Jennifer A. Marshall Graves (The Australian National University, Research School of Biology, The Australian National University, Canberra, Australia); Eric D. Green (National Human Genome Research Institute, National Institutes of Health, Bldg. 50, Rm. 5222, Bethesda, MD); Richard E. Green (Max-Planck Institute for Evolutionary Anthropology, Leipzig Germany); Shannon Hackett (Field Museum of Natural History, Department of Zoology, Division of Birds, Chicago, IL); Paul Hebert (Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada); Kristofer M. Helgen (National Museum of Natural History, MRC 108, Smithsonian Institution, Division of Mammals, PO Box 37012, Washington, DC); Leo Joseph (CSIRO Sustainable Ecosystems—Gungahlin Homestead, Crace ACT 2911, GPO Box 284, Canberra, ACT, Australia); Bailey Kessing (SAIC-Frederick, Inc., National Cancer Institute at Frederick, Laboratory of Genomic Diversity, PO Box B, Frederick, MD); David M. Kingsley (HHMI and Stanford University, Beckman Center B300, Stanford, CA); Harris A. Lewin (Department of Animal Sciences and Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL); Gordon Luikart (Division of Biological Sciences, University of Montana, CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Portugal, Missoula, MT); Paolo Martelli (Ocean Park, Aberdeen, Hong Kong); Miguel A.M. Moreira (Genetics Division, Instituto Nacional de Cancer, Rua Andre Cavalcante, 37, 4o andar, Rio de Janeiro, RJ, Brazil); Ngan Nguyen (University of California, Santa Cruz, Santa Cruz, CA); Guillermo Ortí (George Washington University, Department of Biological Sciences, 2023 G Street, NW, Washington, D.C. 20052); Brian L. Pike (Global Viral Forecasting Initiative, One Market, Spear Tower, Suite 3574, San Francisco, CA); David Michael Rawson (LIRANS Institute, University of Bedfordshire, 250 Butterfield, Great Marlings, Luton, Bedfordshire, UK); Stephan C. Schuster (Penn State University, 310 Wartik Laboratories, University Park, PA); Héctor N. Seuánez (Genetics Division, Instituto Nacional de Câncer, Universidade Federal do Rio de Janeiro, Rua André Cavalcanti 37, 4o andar, Rio de Janeiro, RJ 20231-050, and Department of Genetics, Universidade Federal do Rio de Janeiro, Cidade Universitária, CCS, Bloco A, Rio de Janeiro, RJ 21949-570, Brazil); H. Bradley Shaffer (Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, CA); Mark S. Springer (Department of Biology, University of California, Riverside, CA); Joshua Michael Stuart (University of California, Santa Cruz, Mail Stop SOE2, Santa Cruz, CA); Joanna Sumner (Museum Victoria, GPO Box 666, Melbourne, Vic., Australia); Emma Teeling (University College Dublin, School of Biology and Environmental Science, Science Centre West, Belfield, Dublin, Ireland); Robert C. Vrijenhoek (Monterey Bay Aquarium Research Institute, Moss Landing, CA); Robert D. Ward (CSIRO Marine and Atmospheric Research, GPO Box 1538, Castray Esplanade, Hobart, Tasmania, Australia); Wesley C. Warren (Genome Sequencing Center, Washington University School of Medicine, St Louis, MO); Robert Wayne (UCLA, Ecology

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 Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 32 Jiaochagndong ST, Kunming, Yunnan, China)

Committees:

Mammals Group Members: C. Scott Baker, James Estes, Philippe Gaubert, Jennifer Graves, Alexander Graphodatsky, Kristofer M. Helgen, *****Warren E. Johnson, Harris A. Lewin, Gordon Luikart, *****William J. Murphy, Stephen J. O'Brien, Oliver A. Ryder, Mark Springer, Emma Teeling, Robert Wayne, Terrie Williams, Nathan Wolfe, Ya-Ping Zhang

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

**Amphibians and Reptiles Group Members: ***Andrew J. Crawford, Paolo Martelli, *****Jimmy A. McGuire, *****Robert W. Murphy, H. Bradley Shaffer, *****Barry Sinervo

Fishes Group Members: Fred W. Allendorf, Giacomo Bernardi, Guillermo Orti, David M. Rawson, *****Byrappa Venkatesh, Robert C. Vrijenhoek, Robert D. Ward, *****Edward O. Wiley

General Policy Group Members: C. Scott Baker, *****Adam Felsenfeld, Eric D. Green, *****Robert Hanner, *****Olivier Hanotte, David Haussler, Paul Hebert, Stephen J. O'Brien, Oliver A. Ryder, Hector N. Seuanez, Ya-Ping Zhang

**Analysis Group Members: ***Michele Clamp, Mark Diekhans, David Haussler, Bailey Kessing, David M. Kingsley, Harris A. Lewin, *****Webb Miller, Ngan Nguyen, Brian L. Pike, Stephan C. Schuster, Joshua M. Stuart, Steve Turner

***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/BIA-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

AmphibiaWeb . 2009. Information on amphibian biology and conservation. Available from:http://amphibiaweb.org.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science. 297:1301–1310.

Baker CV. 2008. The evolution and elaboration of vertebrate neural crest cells. Curr Opin Genet Dev. 18:536–543.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science. 304:1321–1325.

Bell MA, Aguirre WE, Buck NJ. 2004. Twelve years of contemporary armor evolution in a threespine stickleback population. Evolution. 58:814–824.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2008. GenBank. Nucleic Acids Res. 36:D25–D30.

Benton MJ, Donoghue PC. 2007. Paleontological evidence to date the tree of life. Mol Biol Evol. 24:26–53.

Blackburn DG. 2000. Reptilian viviparity: past research, future directions, and appropriate models. Comp Biochem Physiol A Mol Integr Physiol. 127:391–409.

Bogart JP, Bi K, Fu J, Noble DW, Niedzwiecki J. 2007. Unisexual salamanders (genus *Ambystoma*) present a new reproductive mode for eukaryotes. Genome. 50:119–136.

Borisenko AV, Sones JE, Hebert PDN. 2009. The front-end logistics of DNA barcoding: challenges and prospects. Mol Ecol Resour. 9: 27–34.

Bossuyt F, Milinkovitch MC. 2000. Convergent adaptive radiations in Madagascan and Asian ranid frogs reveal covariation between larval and adult traits. Proc Natl Acad Sci U S A. 97:6585–6590.

Brito P, Edwards SV. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica. 135:439–455.

Chiari Y, Vences M, Vieites DR, Rabemananjara F, Bora P, Ramilijaona Ravoahangimalala O, Meyer A. 2004. New evidence for parallel evolution of colour patterns in Malagasy poison frogs (Mantella). Mol Ecol. 13:3763–3774.

Collins FS, Morgan M, Patrinos A. 2003. The human genome project: lessons from large-scale biology. Science. 300:286–290.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science. 298:2157–2167.

Duellman WE, Trueb L. 1986. Biology of amphibians. New York: McGraw Hill. 670 pp.

Edwards S, Birks S, Brumfield R. 2005. Future of avian genetic resources collections: archives of evolutionary and environmental history. Auk. 122:979–1284.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. Science. 323:133–138.

Eizirik E, Yuhki N, Johnson WE, Menotti-Raymond M, Hannah SS, O'Brien SJ. 2003. Molecular genetics and evolution of melanism in the cat family. Curr Biol. 13:448–453.

Erwin TL. 1991. How many species are there—revisited. Conserv Biol. 5:330–333.

Eschmeyer WN. 1998. Catalog of fishes. San Francisco (CA): California Academy of Sciences 3 v. (p. 2905)

FAO . 2008. FAO fisheries and aquaculture report. Rome (Italy): Food and Agriculture Organization of the United Nations.

Field D. 2008. Working together to put molecules on the map. Nature. 453:978.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. 2008. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 26:541–547.

Fry BG, Vidal N, Norman JA, Vonk FJ, Scheib H, Ramjan SF, Kuruppu S, Fung K, Hedges SB, Richardson MK, et al. 2006. Early evolution of the venom system in lizards and snakes. Nature. 439:584–588.

Gamble T, Bauer AM, Greenbaum W, Jackman TR. 2008. Out of the blue: a novel, trans-Atlantic clade of geckos (Gekkota, Squamata). Zool Scr. 37:355–366.

Gaston KJ. 1991. The magnitude of global insect species richness. Conserv Biol. 5:283–296.

genome.gov. 2003. Reaffirmation and extension of NHGRI rapid data release policies: large-scale sequencing and other community resource projects [cited 2009 June 28]. Available from:http://www.genome.gov/10506537.

Gregory TR. 2009. Animal genome size database. [cited 2009 May 1]. Available from:http://www.genome.com.

Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, et al. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320:1763–1768.

Hanken J, Jennings DH, Olsson L. 1997. Mechanistic basis of life-history evolution in anuran amphibians: direct development. Am Zool. 37:160–171.

Hanner RH, Gregory TR. 2007. Genomic diversity research and the role of biorepositories. Cell Preserv Technol. 5:93–103.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003. Biological identifications through DNA barcodes. Proc R Soc Lond Series B-Biological Sciences. 270:313–321.

Hedges SB, Couloux A, Vidal N. 2009. Molecular phylogeny, classification, and biogeography of West Indian racer snakes of the tribe Alsophiini (squamata, dipsadidae, xenodontinae). Zootaxa. 2067:1–28.

Hedges SB, Kumar S. 2009. The Timetree of life. New York: Oxford University Press. 572 pp.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695–716.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 61:995–1016.

Holland LZ, Albalat R, Azumi K, Benito-Gutierrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, et al. 2008. The *Amphioxus* genome illuminates vertebrate origins and cephalochordate biology. Genome Res. 18:1100–1111.

Howard R, Moore A, Dickinson EC. 2003. The Howard and Moore complete checklist of the birds of the world. London: Christopher Helm p. 1039.

James TY, Litvintseva AP, Vilgalys R, Morgan JA, Taylor JW, Fisher MC, Berger L, Weldon C, du Preez L, Longcore JE. 2009. Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. PLoS Pathog. 5:e1000458.

Janvier P. 2006. Palaeontology: modern look for ancient lamprey. Nature. 443:921–924.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science. 188:107–116.

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK. 2006. Genomics and conservation genetics. Trends Ecol Evol. 21:629–637.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Liu H, Zhu F, Yong J, Zhang P, Hou P, Li H, Jiang W, Cai J, Liu M, Cui K, et al. 2008. Generation of induced pluripotent stem cells from adult rhesus monkey fibroblasts. Cell Stem Cell. 3:587–590.

Maisey JG. 2000. Discovering fossil fishes. Boulder (CA): Westview Press. 224 pp.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–141.

Marjanovic D, Laurin M. 2007. Fossils, molecules, divergence times, and the origin of lissamphibians. Syst Biol. 56:369–388.

May RM. 1988. How many species are there on Earth. Science. 241: 1441–1449.

Mendelson JR 3rd, Lips KR, Gagliardo RW, Rabb GB, Collins JP, Diffendorfer JE, Daszak P, Ibanez DR, Zippel KC, Lawson DP, et al. 2006. Biodiversity. Confronting amphibian declines and extinctions. Science. 313:48.

Meulemans D, Bronner-Fraser M. 2002. *Amphioxus* and lamprey AP-2 genes: implications for neural crest evolution and migration patterns. Development. 129:4953–4962.

Murphy RW, Fu J, Darevsky IS, Kupriyanova LA, MacCulloch RD. 2000. A fine line between sex and unisexuality: the phylogenetic constraints on parthenogenetic lacertid lizards. Zool J Linn Soc. 130:527–549.

Nachman MW, Hoekstra HE, D'Agostino SL. 2003. The genetic basis of adaptive melanism in pocket mice. Proc Natl Acad Sci U S A. 100: 5268–5273.

Nowak RM. 1999. Walker's mammals of the world. 6th ed. Baltimore (MD): Johns Hopkins University Press. 1936 pp.

O'Brien SJ. 2003. Tears of the cheetah: and other tales from the genetic frontier. 1st ed. New York,: Thomas Dunne Books/St. Martin's Press. xiv, p. 287.

Okita K, Ichisaka T, Yamanaka S. 2007. Generation of germline-competent induced pluripotent stem cells. Nature. 448:313–317.

Osorio J, Retaux S. 2008. The lamprey in evolutionary studies. Dev Genes Evol. 218:221–235.

Parham JW, Bickham JW, Iverson JB, Philippen HD, Rhodin AGJ, Shaffer HB, Spinks PQ, van Dijk PP. 2007. An annotated list of modern turtle terminal taxa, with comments on areas of taxonomic instability and recent change. In: 4 CRMN, (ed) Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises. pp173–1994.

Ratnasingham S, Hebert PD. 2007. bold: the barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes. 7:355–364.

Romanov MN, Tuttle EM, Houck ML, Modi WS, Chemnick LG, Korody ML, Mork EM, Otten CA, Renner T, Jones KC, et al. 2009. The value of avian genomics to the conservation of wildlife. BMC Genomics. 10(Suppl. 2):S10.

Ryder OA. 2005. Conservation genomics: applying whole genome studies to species conservation efforts. Cytogenet Genome Res. 108:6–15.

Ryder OA, McLaren A, Brenner S, Zhang YP, Benirschke K. 2000. DNA banks for endangered animal species. Science. 288:275–277.

Sanderson MJ. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol. 14: 1218–1231.

Sansom IJ, Smith MM, Smith MP. 1996. Scales of thelodont and shark-like fishes from the Ordovician of Colorado. Nature. 379:628–630.

Schwartz M, Luikart G, McKelvey K, Cushman S. 2009. Landscape genomics: a brief perspective. In: Huettmann F, Cushman S, editors. Spatial complexity, informatics and wildlife conservation. Tokyo (Japan): Springer.

Shapiro MD, Bell MA, Kingsley DM. 2006. Parallel genetic origins of pelvic reduction in vertebrates. Proc Natl Acad Sci U S A. 103:13753–13758.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotechnol. 26:1135–1145.

Shimeld SM, Holland PW. 2000. Vertebrate innovations. Proc Natl Acad Sci U S A. 97:4449–4452.

Shubin N, Wake DB, Crawford AJ. 1995. Morphological variation in the limbs of *taricha-granulosa (caudata, salamandridae)*—evolutionary and phylogenetic implications. Evolution. 49:874–884.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou MM, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, Angerer LM, Arnone MI, Burgess DR, Burke RD, et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science. 314:941–952.

Stadtfeld M, Nagaya M, Utikal J, Weir G, Hochedlinger K. 2008. Induced pluripotent stem cells generated without viral integration. Science. 322:945–949.

Stiassny M, Wiley E, Johnson G. 2004. Gnathostome fishes. In: Cracraft J, Donoghue M, editors. Assembling the tree of life. New York: Oxford University Press. p. 410–429.

Stockwell CA, Hendry AP, Kinnison MT. 2003. Contemporary evolution meets conservation biology. Trends Ecol Evol. 18:94–101.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. 2007. Population genomics of human gene expression. Nat Genet. 39:1217–1224.

Thomas G, Quoss H, Hartmann J, Eckmann R. 2009. Human-induced changes in the reproductive traits of Lake Constance common whitefish (*Coregonus lavaretus*). J Evol Biol. 22:88–96.

Tidwell JH, Allan GL. 2001. Fish as food: aquaculture's contribution. Ecological and economic impacts and contributions of fish farming and capture fisheries. EMBO Rep. 2:958–963.

Trounson A. 2009. Rats, cats, and elephants, but still no unicorn: induced pluripotent stem cells from new species. Cell Stem Cell. 4:3–4.

Uetz P. 2009. The TIGR Reptile Database. Available from:http://www.reptile-database.org/.

Uetz P, Goll J, Hallermann J. 2007. Die tigr-reptiliendatenbank. Elaphe. 15:16–19.

van Tuinen M, Stidham M, Hadly TA. 2006. Tempo and mode of modern bird evolution observed with large-scale taxonomic sampling. Historical Biol. 18:205–221.

Vanhoye D, Bruston F, Nicolas P, Amiche M. 2003. Antimicrobial peptides from hylid and ranin frogs originated from a 150-million-year-old ancestral precursor with a conserved signal peptide but a hypermutable antimicrobial domain. Eur J Biochem. 270:2068–2081.

Vences M, Kosuch J, Boistel R, Haddad CFB, La Marca E, Lotters S, Veith M. 2003. Convergent evolution of aposematic coloration in Neotropical poison frogs: a molecular phylogenetic perspective. Org Divers Evol. 3:215–226.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. Science. 291:1304–1351.

Vidal N, Hedges SB. 2009. The molecular evolutionary tree of lizards, snakes, and amphisbaenians. C R Biol. 332:129–139.

Wang Z, Miyake T, Edwards SV, Amemiya CT. 2006. Tuatara (Sphenodon) genomics: bAC library construction, sequence survey, and application to the DMRT gene family. J Hered. 97:541–548.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36:D13–D21.

Wilson DE, Reeder DM. 2005. Mammal species of the world: a taxonomic and geographic reference. 3rd ed. Baltimore (MD): Johns Hopkins University Press. 2 v. (xxxv, p. 2142).

Yu J. 2009. Human induced pluripotent stem cells free of vector and transgene sequences. Science. 324:797–801.

Yusa K, Rad R, Takeda J, Bradley A. 2009. Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. Nat Methods. 6:363–369.

Zaher H, Grazziotin FG, Cadle JE, Murphy RW, Moura-Leite JCd, Bonatto SL. 2009. Molecular phylogeny of advanced snakes (Serpentes, Caenophidia) with an emphasis on South American xenodontines: a revised classification and descriptions of new taxa. Pap Avuls Zool. 49: 115–153.

Zasloff M. 2002. Antimicrobial peptides of multicellular organisms. Nature. 415:389–395.

ZSL . 2009. EDGE: evolutionary distinct & globally endangered. The Zoological Society of London Available from:http://www.edgeofexistence.org/index.php

Received July 22, 2009; Revised September 21, 2009; Accepted September 22, 2009

Corresponding Editor: Stephen Karl

LGDMSNO09700

October 29, 2009

Journal of Heredity

Supplemental Materials

Appendix 1

Policy issues

- The G10K consortium members recognize this endeavor to be an open access, noncommercial research program without “reach-through” intellectual property (IP) agreements.
- All individual(s)/institute(s) contributing biomaterials for sequencing thus comply with the above mentioned noncommercial policy and acknowledge that contributed samples have been collected and distributed in accordance with applicable national and international laws and regulations.
- The reference materials used for genome sequencing should be clearly identified, accessible and where possible, adhere to best practices for biodiversity repositories (particularly concerning taxonomic identification of voucher specimens/tissues and genome size estimation (Hanner and Gregory 2007)). Reference documentation should include digital images/CT scans of the morphological voucher specimens, geospatial coordinates of the specimen collection site (Field 2008) and other emerging metadata standards proposed for genome sequences (Field et al., 2008).
- Deposition and publication will follow the Ft Lauderdale agreement (genome.gov 2003) in accord with prior best practices in genomic research, including release of

primary sequence data to the international databases within 24 hours and release of the full genome assemblies as soon as quality assurance is complete.

- The consortium is committed to building programs for training of highly qualified personnel, particularly in developing nations.

Appendix 2

Detailed table of classes, orders, families, genera and species

Groups	Orders	Families			Genera			Species		
		G10K samples available	Total	% of Total	G10K samples available	Total	% of Total	G10K samples available	Total	% of Total
Mammals	Afrosoricida	2	2	100%	12	19	63%	18	51	35%
Mammals	Carnivora	15	15	100%	98	126	77%	163	287	56%
Mammals	Cetartiodactyla	21	21	100%	120	129	93%	238	324	73%
Mammals	Chiroptera	18	18	100%	115	203	57%	316	1116	28%
Mammals	Cingulata	1	1	100%	4	9	44%	5	21	23%
Mammals	Dasyuromorphia	2	2	100%	19	23	83%	55	71	77%
Mammals	Dermoptera	1	1	100%	1	2	50%	1	2	50%
Mammals	Didelphimorphia	1	1	100%	12	17	70%	34	87	39%
Mammals	Diprotodontia	11	11	100%	34	39	87%	88	143	61%
Mammals	Eulipotyphla	4	4	100%	28	55	51%	54	452	11%
Mammals	Hyracoidea	1	1	100%	3	3	100%	4	4	100%
Mammals	Lagomorpha	2	2	100%	7	12	58%	21	91	23%
Mammals	Macroscelidea	1	1	100%	3	4	75%	4	15	26%
Mammals	Microbiotheria	1	1	100%	1	1	100%	1	1	100%
Mammals	Monotremata	2	2	100%	3	3	100%	4	5	80%
Mammals	Notoryctemorphia	1	1	100%	1	1	100%	2	2	100%
Mammals	Paucituberculata	1	1	100%	3	3	100%	3	6	50%
Mammals	Peramelemorphia	2	3	66%	6	8	75%	12	21	57%
Mammals	Perissodactyla	3	3	100%	6	6	100%	16	17	94%
Mammals	Pholidota	1	1	100%	1	1	100%	6	8	75%
Mammals	Pilosa	4	4	100%	5	5	100%	8	10	80%
Mammals	Primates	14	15	93%	54	69	78%	146	376	38%
Mammals	Proboscidea	1	1	100%	2	2	100%	3	3	100%
Mammals	Rodentia	30	33	90%	219	481	45%	612	2277	26%
Mammals	Scandentia	2	2	100%	3	5	60%	9	20	45%
Mammals	Sirenia	2	2	100%	2	3	66%	2	5	40%
Mammals	Tubulidentata	1	1	100%	1	1	100%	1	1	100%
Mammals Count	**27**	**145**	**150**	**97%**	**763**	**1230**	**62%**	**1826**	**5416**	**34%**
Birds	Accipitriformes	2	2	100%	51	72	71%	121	240	50%
Birds	Aegothelidae	1	1	100%	1	2	50%	2	9	22%
Birds	Anseriformes	3	3	100%	41	52	78%	130	162	80%

Groups	Orders	Families			Genera			Species		
		G10K samples available	Total	% of Total	G10K samples available	Total	% of Total	G10K samples available	Total	% of Total
Birds	Apodiformes	3	3	100%	99	125	79%	241	429	56%
Birds	Caprimulgiformes (Part 1, 2,3)	4	4	100%	16	20	80%	47	109	43%
Birds	Cariamidae	1	1	100%	1	2	50%	1	2	50%
Birds	Charadriiformes	14	17	82%	70	88	80	226	367	62%
Birds	Ciconiiformes	3	3	100%	32	39	82%	72	116	62%
Birds	Coliiformes	1	1	100%	2	2	100%	5	6	83%
Birds	Columbiformes	1	2	50%	31	44	70%	136	311	43%
Birds	Coraciiformes (Part 1, 2)	9	10	90%	36	50	72%	103	208	50%
Birds	Cuculiformes	1	1	100%	22	36	61%	57	138	41%
Birds	Falconiformes	1	1	100%	9	11	82%	40	64	63%
Birds	Galliformes	5	5	100%	58	80	72%	135	290	46%
Birds	Gaviiformes	1	1	100%	1	1	100%	3	5	60%
Birds	Gruiformes	7	7	100%	27	44	61%	72	165	44%
Birds	Leptosomatidae	0	1	0%	0	1	0%	0	1	0%
Birds	Mesitomithiformes	0	1	0%	0	2	0%	0	3	0%
Birds	Musophagiformes	1	1	100%	5	6	83%	16	23	69%
Birds	Opisthicomiformes	1	1	100%	1	1	100%	1	1	100%
Birds	Otididae	1	1	100%	4	11	36%	6	26	23%
Birds	Paleognathae	6	6	100%	14	15	93%	37	59	62%
Birds	Passeriformes	91	101	90%	873	1226	71%	2951	5756	51%
Birds	Pelecaniformes	7	7	100%	8	9	88%	34	62	54%
Birds	Phaethontiformes	1	1	100%	1	1	100%	3	3	100%
Birds	Phoenicopteriformes	1	1	100%	3	3	100%	4	5	80%
Birds	Piciformes	5	5	100%	53	68	77%	255	398	64%
Birds	Podicipediformes	1	1	100%	6	6	100%	13	22	59%
Birds	Procellariiformes	4	4	100%	20	27	74%	56	112	50%
Birds	Psittaciformes	1	1	100%	69	85	81%	201	364	55%
Birds	Pteroclideformes	1	1	100%	1	2	50%	2	16	13%
Birds	Sphenisciformes	1	1	100%	6	6	100%	13	17	76%
Birds	Strigiformes	2	2	100%	21	29	72%	66	195	33%
Birds	Trogoniformes	1	1	100%	5	6	83%	25	39	64%
Birds Count	**34**	**182**	**199**	**91%**	**1587**	**2172**	**73%**	**5074**	**9723**	**52%**

Groups	Orders	Families			Genera			Species		
		G10K samples available	Total	% of Total	G10K samples available	Total	% of Total	G10K samples available	Total	% of Total
Amphibians	Anura	35	40	87%	248	408	60%	1609	5811	27%
Amphibians	Caudata	10	10	100%	40	69	57%	133	583	22%
Amphibians	Gymnophiona	5	6	83%	13	33	39%	18	176	10%
Amphibians Count	**3**	**50**	**56**	**89%**	**301**	**510**	**59%**	**1760**	**6570**	**27%**
Reptiles	Crocodylia	3	3	100%	8	9	88%	20	23	86%
Reptiles	Rhynchocephalia	1	1	100%	1	1	100%	1	2	50%
Reptiles	Squamata	47	48	97%	672	983	68%	3136	8664	36%
Reptiles	Testudines	12	13	92%	70	94	74%	140	313	44%
Nonavian Reptiles Count	**4**	**63**	**65**	**97%**	**751**	**1087**	**69%**	**3297**	**9002**	**37%**
Fishes	Acipenseriformes	2	2	100%	4	6	66%	8	28	28%
Fishes	Albuliformes	1	1	100%	1	2	50%	2	11	18%
Fishes	Amiiformes	1	1	100%	1	1	100%	1	1	100%
Fishes	Anguilliformes	13	15	86%	40	153	26%	92	903	10%
Fishes	Ateleopodiformes	1	1	100%	2	4	50%	2	13	15%
Fishes	Atheriniformes	7	10	70%	19	51	37%	54	326	16%
Fishes	Aulopiformes	13	16	81%	27	44	61%	60	256	23%
Fishes	Batrachoidiformes	1	1	100%	6	23	26%	10	80	12%
Fishes	Beloniformes	5	5	100%	21	34	61%	55	244	22%
Fishes	Beryciformes	7	7	100%	18	30	60%	52	158	32%
Fishes	Carcharhiniformes	6	8	75%	32	50	64%	96	277	34%
Fishes	Ceratodontiformes	1	1	100%	1	1	100%	1	1	100%
Fishes	Cetomimiformes	1	2	50%	5	18	27%	7	33	21%
Fishes	Characiformes	14	18	77%	85	278	30%	184	1898	9%
Fishes	Chimaeriformes	3	3	100%	5	6	83%	17	46	36%
Fishes	Clupeiformes	4	6	66%	24	84	28%	48	393	12%
Fishes	Coelacanthiformes	1	1	100%	1	1	100%	1	2	50%
Fishes	Cypriniformes	4	6	66%	47	451	10%	96	3943	2%
Fishes	Cyprinodontiformes	9	10	90%	33	118	27%	101	1172	8%
Fishes	Elopiformes	2	2	100%	2	2	100%	5	8	62%
Fishes	Esociformes	2	2	100%	4	4	100%	6	12	50%
Fishes	Gadiformes	9	10	90%	49	81	60%	107	610	17%
Fishes	Gasterosteiformes	5	5	100%	10	11	90%	11	29	37%
Fishes	Gobiesociformes	1	1	100%	4	46	8%	6	154	3%
Fishes	Gonorynchiformes	2	4	50%	2	7	28%	2	37	5%
Fishes	Gymnotiformes	5	5	100%	8	32	25%	11	158	6%

Groups	Orders	Families			Genera			Species		
		G10K samples available	Total	% of Total	G10K samples available	Total	% of Total	G10K samples available	Total	% of Total
Fishes	Heterodontiformes	1	1	100%	1	1	100%	4	9	44%
Fishes	Hexanchiformes	2	2	100%	4	4	100%	5	5	100%
Fishes	Lamniformes	6	7	85%	8	10	80%	11	17	64%
Fishes	Lampriformes	6	7	85%	10	12	83%	12	27	44%
Fishes	Lepidosireniformes	2	2	100%	2	2	100%	2	5	40%
Fishes	Lepisosteiformes	1	1	100%	2	2	100%	4	7	57%
Fishes	Lophiiformes	12	18	66%	25	67	37%	41	328	12%
Fishes	Myctophiformes	2	2	100%	18	36	50%	37	262	14%
Fishes	Myxiniformes	1	1	100%	2	5	40%	4	71	5%
Fishes	Notacanthiformes	2	2	100%	5	6	83%	9	26	34%
Fishes	Ophidiiformes	3	5	60%	25	116	21%	43	501	8%
Fishes	Orectolobiformes	6	7	85%	9	13	69%	20	42	47%
Fishes	Osmeriformes	11	13	84%	37	77	48%	68	315	21%
Fishes	Osteoglossiformes	7	7	100%	10	32	31%	11	219	5%
Fishes	Perciformes	129	164	78%	684	1722	39%	1892	10663	17%
Fishes	Percopsiformes	3	3	100%	4	7	57%	5	9	55%
Fishes	Petromyzontiformes	1	1	100%	4	8	50%	5	43	11%
Fishes	Pleuronectiformes	11	11	100%	80	136	58%	152	793	19%
Fishes	Polymixiiformes	1	1	100%	1	1	100%	3	10	30%
Fishes	Polypteriformes	1	1	100%	1	2	50%	1	12	8%
Fishes	Pristiformes	1	1	100%	2	2	100%	4	7	57%
Fishes	Pristiophoriformes	1	1	100%	1	2	50%	2	6	33%
Fishes	Rajiformes	12	12	100%	43	70	61%	143	574	24%
Fishes	Saccopharyngiformes	4	4	100%	4	5	80%	5	28	17%
Fishes	Salmoniformes	1	1	100%	6	11	54%	12	210	5%
Fishes	Scorpaeniformes	25	35	71%	107	294	36%	257	1573	16%
Fishes	Siluriformes	18	38	47%	68	478	14%	125	3400	3%
Fishes	Squaliformes	4	4	100%	12	23	52%	40	126	31%
Fishes	Squatiniformes	1	1	100%	1	1	100%	6	22	27%
Fishes	Stephanoberyciformes	2	4	50%	5	11	45%	6	55	10%
Fishes	Stomiiformes	4	4	100%	38	53	71%	69	423	16%
Fishes	Synbranchiformes	2	3	66%	5	13	38%	7	110	6%
Fishes	Syngnathiformes	4	5	80%	18	64	28%	39	336	11%
Fishes	Tetraodontiformes	10	10	100%	72	104	69%	143	436	32%
Fishes	Torpediniformes	2	4	50%	2	12	16%	8	68	11%
Fishes	Zeiformes	5	6	83%	10	16	62%	16	33	48%

Groups	Orders	Families			Genera			Species		
		G10K samples available	Total	% of Total	G10K samples available	Total	% of Total	G10K samples available	Total	% of Total
Fishes Count	**62**	**424**	**532**	**80%**	**1777**	**4956**	**36%**	**4246**	**31564**	**13%**

Appendix 3

Data Analysis

The G10K project intends to produce assembled whole chromosomes with high enough quality to support bioinformatics analysis, including whole-genome multiple alignment, determination of lines of descent for segments of DNA with sizes ranging from single bases to multi-mega-base chromosomal regions, as well as gene finding and the identification of other functional elements by patterns of selection. This will require the coordination of a network of computational centers to assemble, align, store, and disseminate via on-line genome browsers the vast amount of sequencing and annotation information. A central sample tracking database would also track samples and their quality and link taxonomic data with important phenotypic attributes (Appendix 4). A cloud computing infrastructure could support the availability of algorithms for efficient large-scale analysis of these data.

To put the ambition of G10K into perspective, we present the following test calculation for the sample throughput and processing requirements for the sequencing and assembly stages. A five-year project requires 2000 genomes to be sequenced and assembled every year. It seems reasonable to distribute the workload over 20 sites, resulting in 100 samples (genomes) to be completed by one sequencing site per year. This translates to a required output of two genomes per week for each sequencing site, including both sequencing and draft assembly.

Distributing and sequencing 10,000 sequence-ready samples across 20 sequencing sites at a rate of 100 per site per year for five years necessitates a highly structured workflow that minimizes delays in sequencing. We foresee a distributed network of sequencing facilities, each represented by a coordinator that will communicate with a large data center (Figure 1 of Appendix 3).

Assembly of vertebrate genomes using new sequencing technologies is still an active research topic. While we anticipate that assembly methods will improve in the next few years, with current technology a full assembly requires the use of a large memory machine (~1.5Tb RAM), 128 CPUs and 2 weeks to complete. It is likely that new sequencing technologies will take advantage of some derivative of cloud computing technology, or equivalent large scale distributed computing resources, for assembly and initial analysis of data from a sequencing run. This activity could be coordinated remotely by the data center. A primary driver here is the fact that the amount of storage for a vertebrate sequencing run can reach many Tb of disk. To transfer all raw data to a central location daily would exceed the capacity of most networks; therefore, initial steps in analysis of each sequencing run must be conducted in a distributed fashion.

We envision the output of the initial analysis of a sequencing run to take the form of a ‘minimal’ assembly of all reads into (possibly small) contigs. The central data center will then download this assembly in a format taking up a much smaller amount of storage space. For example, the files could be communicated in FASTA-formatted sequences with quality scores (fastq format), or some more expressive variant. This ‘minimal’ assembly will be archived, accessioned and made publically available. The current protocol for data release is that every contig over 1000 bases is made public. As the precise nature of the data release depends on the properties of the technology used, the spirit of timely public data release should be maintained.

The completeness of each species’ assembly will depend on the sequencing technology employed. The most useful assembly for annotation and comparative genomics is a “complete” assembly, which includes deep enough coverage for whole-chromosome assemblies, i.e., sequence coverage that is sufficient to (ideally) create

one sequence contig per chromosome, and identify DNA polymorphisms within and between species. This is a seminal goal of the project.

Analysis and Annotation

Initial computational analysis, annotation and visualization of the Genome 10K data is a very open-ended aspect of the project, but can be anticipated to demand at least as much computational effort as sequence assembly. Most of this will be performed at the data center. It will critically rely on whole-genome multiple alignments, currently one of the most computation-intensive steps in vertebrate genome analysis, requiring large clusters of machines. It is unclear how difficult this problem will be, as whole-genome multiple alignments for more than 50 vertebrate genomes have never been built. We suggest a computational pipeline for analysis, annotation and visualization be specified soon, and put in place and tested prior to the start of sequencing. This pipeline would serve as an evaluation of assembly, analysis and annotation methods prior to their deployment.

Timing

Due to the anticipated rapid improvement in technology, some decisions should be postponed as late as possible to allow for a better estimate of precise parameters (e.g. exact sample size and assembly method). However, it is imperative that the sample tracking database (Appendix 4) be operational as soon as possible in order to store the thousands of existing samples.

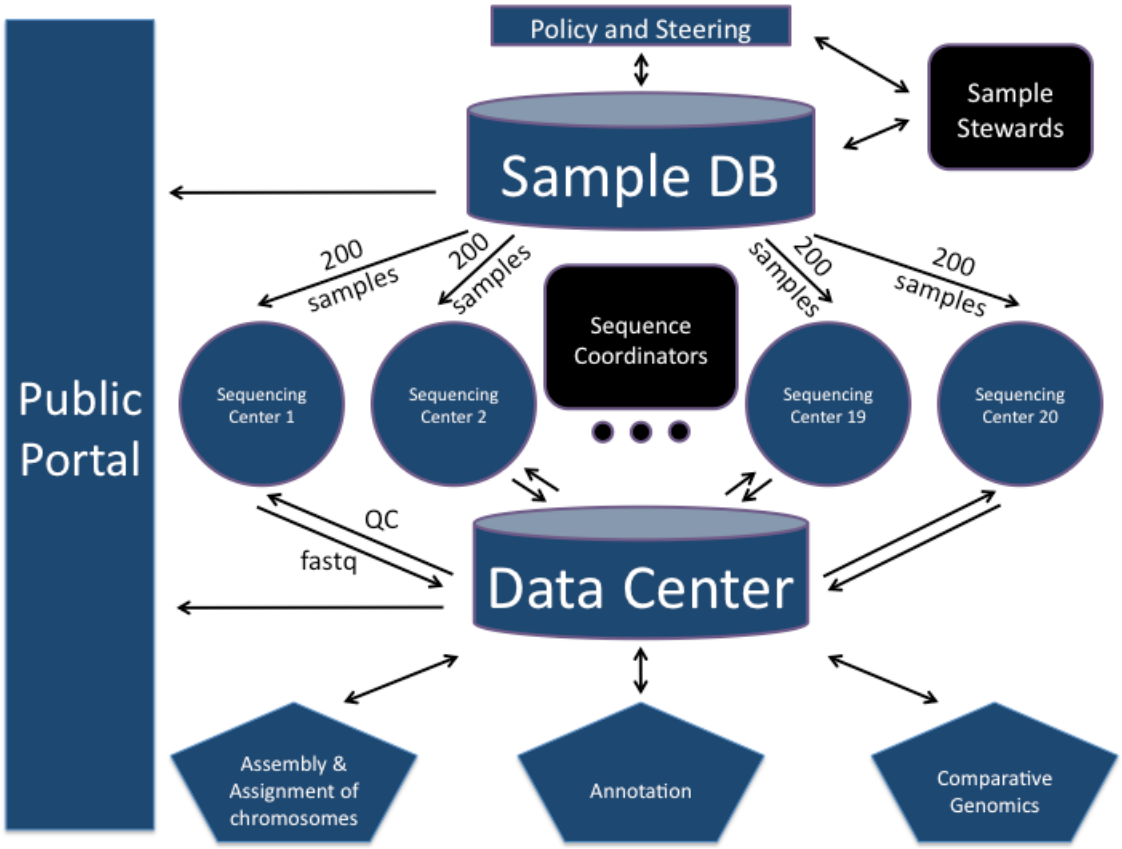

Figure 1 of Appendix 3

This is an overview of the sample and data flow through the multiple sequencing centers. A central tracking database (Sample DB) will store the progress of all species through the sequencing and analysis process and will be publicly available.

Appendix 4

Genome 10K Sample Tracking Database Design

1. Introduction

This document describes the design of the Genome 10K sample tracking database. The goal of this database is to catalog samples available to the Genome 10K program. It will provide sufficient information to allow the project to prioritize and select samples for sequencing. Tracking of samples during the sequencing and assembly processes will be supported.

2. Taxonomies

Tracking taxonomy for animals is the key to evaluating the state of the Genome 10K project and to prioritize obtaining and sequencing samples. This is greatly complicated by the fact that there is no one agreed upon taxonomy covering all vertebrates. Different clade-specific taxonomic trees are preferred by experts in each of the species groups. While the NCBI taxonomy (Benson et al., 2008; Wheeler et al., 2008) covers all vertebrates, it often disagrees with the preferred clade-specific trees. Mapping between the taxonomies is not straightforward and is often not possible due to conflicting classifications or the incompleteness of taxonomic branches.

The Genome 10K project will not attempt to dictate a particular taxonomy for the entire project. Instead groups may use clade-specific taxonomies for their samples and these will be recorded in the database. A G10K hybrid taxonomy will be constructed for reporting purposes by attaching the clade-specific trees to a vertebrate wide taxonomy tree “backbone”. Species will also be mapped to the leaves of alternate taxonomy trees when these are available. This will permit different analyses to use different trees.

Taxonomies are not static and undergo periodic updates. The NCBI taxonomy database is updated daily, while others undergo slower rates of change. The G10K database will store multiple versions of each taxonomy, supporting migration to new versions. We will not attempt to track every version of NCBI's taxonomy; it will be updated periodically or as needed by the project.

The following taxonomies will be imported into the Genome 10K database if possible:

- NCBI taxonomy database (Benson et al., 2008; Wheeler et al., 2008) (http://www.ncbi.nlm.nih.gov/Taxonomy/
- Wilson and Reeder: Mammal Species of the World (Wilson and Reeder 2005) (http://www.bucknell.edu/MSW3/)
- Amphibian Species of the World: an Online Reference (Frost 2009) (http://research.amnh.org/herpetology/amphibia/)
- TIGR Reptile Database (Uetz et al., 2007) (http://www.reptile-database.org/)
- Catalog of Fishes (Eschmeyer 1998) (http://research.calacademy.org/research/Ichthyology/Catalog/fishcatmain.asp)
- Sibley Bird database (Monroe and Sibley 1993; Sibley and Ahlquist 1990; Sibley and Monroe 1993) (http://www.scricciolo.com/classificazione/cover.html)
- Tree of Life project (Maddison and Schulz 2007) (http://tolweb.org)

3. Database organization

This section describes the logical organization of the G10K sample tracking database. It is a high-level description of the data that will be managed for the project. This attempts to be comprehensive while omitting the details of the SQL schema that implements the database.

3.1 Species record

The species records provide the link between the externally defined taxonomy trees and the species names in the database, along with any species-specific data tracked by the project.

The following fields are defined:

- G10K species ID
- class
- order
- family
- genus
- species
- subspecies
- common name
- G10K taxonomy tree - taxonomy tree and version used to define the taxonomy for this species in the G10K project
- species ID in G10K taxonomy tree
- list of other taxonomies and corresponding species IDs for this species in those other taxonomies
- USESA (U.S. Endangered Species Act status
- CITES (Goldsmith 1978) appendix status

3.2 Animal record

An animal record is kept for each individual animal tracked by the database. Biological samples from an animal are tracked in separate records, allowing multiple samples to be tracked for a given individual animal.

The following fields are defined:

- G10K animal ID - ID number assigned by G10K project
- species ID - reference to the G10K species record
- barcode - BOLD (Ratnasingham and Hebert 2007) barcode accession
- sex
- steward – steward for the animal or animal tissues
- steward's animal ID - identifier for animal used by the steward
- date of birth
- location of birth - longitude, latitude, depth
- date of death
- date of collection
- location of collection - longitude, latitude, depth
- collector
- voucher ID number
- voucher location
- steward holds permit for collection (yes, no, unknown, NA)
- ACUC permit status - steward has permit of their institution's Animal Care and Use Committee (confirmed, pending, unknown, NA)
- USESA permit status (confirmed, pending, unknown, NA)
- CITES permit status (confirmed, pending, unknown, NA)

3.3 Sample record

A sample record describes a biological sample from an animal. Multiple samples may be created from the same or different tissues from the same animal, including cell lines.

The following fields are defined:

- G10K sample ID - ID number assigned by G10K project
- animal ID - reference to G10K animal record
- steward – steward for the sample
- steward's sample ID – sample identifier for specimen used by the steward
- tissue from which sample is taken
- sample type (DNA, blood, other tissue, cell line)
- sample quantity
- sample quality
- storage location
- preservative type
- fields to track progress of sample in G10K pipeline (to be specified)

3.4 Steward record

The steward records track the individuals and institutions responsible for each animal and sample in the database. Animals and samples may have different stewards. The following fields are defined:

- G10K steward ID
- name
- institution
- address
- email
- telephone
- fax

References:

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL, 2008. GenBank. Nucleic Acids Res. 36**:** D25-D30.

Eschmeyer WN, 1998. Catalog of fishes. California Academy of Sciences: San Francisco, 3 v. (2905 p.)pp.

Field D, 2008. Working together to put molecules on the map. Nature. 453**:** 978.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV et al., 2008. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 26**:** 541-547.

Frost DR. 2009. Version 5.3 (12 February, 2009). Amphibian species of the world: an online reference. Available from: http://research.amnh.org/herpetology/amphibia/

genome.gov. 2003. Reaffirmation and extension of NHGRI rapid data release policies: large-scale sequencing and other community resource projects. [cited 2009 June 28]. Available from: http://www.genome.gov/10506537.

Goldsmith EI, 1978. The Convention on International Trade in Endangered Species of Wild Fauna and Flora. J Med Primatol. 7**:** 122-124.

Hanner RH, Gregory TR, 2007. Genomic diversity research and the role of biorepositories. Cell Preserv. Technol. 5**:** 93-103.

Maddison DR, Schulz K-S. 2007. The Tree of Life Web Project. Available from: http://tolweb.org.

Monroe BL, Sibley CG, 1993. A world checklist of birds. Yale University Press: New Haven, xix, 393 p.pp.

Ratnasingham S, Hebert PD, 2007. bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes. 7**:** 355-364.

Sibley CG, Ahlquist JE, 1990. Phylogeny and classification of birds : a study in molecular evolution. Yale University Press: New Haven, xxiii, 976 p.pp.

Sibley CG, Monroe BL, 1993. A supplement to distribution and taxonomy of birds of the world. Yale University Press: New Haven, vi, 108 p.pp.

Uetz P, Goll J, Hallermann J, 2007. Die TIGR-Reptiliendatenbank. Elaphe. 15**:** 16-19.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S et al., 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36**:** D13-D21.

Wilson DE, Reeder DM, 2005. Mammal species of the world : a taxonomic and geographic reference. 3rd edn. Johns Hopkins University Press: Baltimore, 2 v. (xxxv, 2142 p.)pp.