Perspectives from the Avian Phylogenomics Project: Questions that Can Be Answered with Sequencing All Genomes of a Vertebrate Class

Erich D. Jarvis

Department of Neurobiology, Duke University Medical Center, Durham, North Carolina 27710
Howard Hughes Medical Institute, Chevy Chase, Maryland 20815

Annu. Rev. Anim. Biosci. 2016. 4:45–59

The *Annual Review of Animal Biosciences* is online at animal.annualreviews.org

This article's doi:
10.1146/annurev-animal-021815-111216

Keywords

B10K, avian genomes, species tree, trait evolution, comparative genomics, extinction, algorithms

Abstract

The rapid pace of advances in genome technology, with concomitant reductions in cost, makes it feasible that one day in our lifetime we will have available extant genomes of entire classes of species, including vertebrates. I recently helped cocoordinate the large-scale Avian Phylogenomics Project, which collected and sequenced genomes of 48 bird species representing most currently classified orders to address a range of questions in phylogenomics and comparative genomics. The consortium was able to answer questions not previously possible with just a few genomes. This success spurred on the creation of a project to sequence the genomes of at least one individual of all extant ~10,500 bird species. The initiation of this project has led us to consider what questions now impossible to answer could be answered with all genomes, and could drive new questions now unimaginable. These include the generation of a highly resolved family tree of extant species, genome-wide association studies across species to identify genetic substrates of many complex traits, redefinition of species and the species concept, reconstruction of the genomes of common ancestors, and generation of new computational tools to address these questions. Here I present visions for the future by posing and answering questions regarding what scientists could potentially do with available genomes of an entire vertebrate class.

INTRODUCTION

With the sequencing of the first draft of the human genome in the early 2000s, there was hope that many new scientific questions could be addressed. However, this was not possible for most other species, as the amount of time and financial investment was enormous to sequence just one human genome [over $2.7 billion (1, 2)]. As sequencing became cheaper, genomes of the most popular animal models, such as mouse (3, 4), chicken (5), and zebra finch (6), were sequenced with faster and cheaper approaches, but still with costs in the tens of millions. At that time, it was not readily imaginable that a scientist could have access to sequenced genomes of 10 or more of the species they use to conduct their studies. However, with the advent of next-generation sequencing beginning in 2004, scientists could foresee a future where genomes of many species could be sequenced and assembled de novo quickly and for a reasonable cost (7). As a result, the Genome 10K (G10K) Consortium was formed in 2009 with the goal of fostering sequencing of 10,000 vertebrate genomes (out of an estimated 66,000 species) within 5–10 years (8, 9). Sequencing companies, such as BGI in China, began ambitious projects to sequence many animal and plant genomes, either as part of their mission or in collaboration with scientists around the world (10). In 2010, BGI offered to collaborate with the G10K Consortium on the sequencing of the first 100 representative vertebrate species across each major vertebrate group, selected by the G10K researchers. One outgrowth of this collaborative effort, combined with others conducting sequencing of bird genomes, was the Avian Phylogenomics Project (11).

The Avian Phylogenomics Project became the most ambitious multispecies vertebrate genome sequencing project between 2010 and 2014. Our main mission was to sequence the genomes of at least one representative species per order and use those genomes to resolve debates on the origin of modern birds; their ordinal relationships and genome evolution; and the genetics of several complex traits, including vocal learning. This effort required bringing together many researchers involved in sequencing avian and other vertebrate genomes (e.g., outgroups), incorporating and developing ongoing improvements to sequencing technology, genome assembly, annotation, and analyses (12, 13). The result was a four-year effort focused on generation or collection of genomes of 48 avian and 3 reptilian species, using human as an outgroup mammal. Some studies used the genomes of 10 or more mammal species from public databases, but these were not as evenly spaced across orders. To date, nearly 50 papers have been published from this effort, with most appearing in special issues of *Science*, *Genome Biology*, and *GigaScience* and others in many other journals (11; **http://avian.genomics.cn/en/**). Some of the highlights of the studies were the generation of a highly resolved genome-scale phylogenomic tree for the early branches in the tree of life of modern birds (13); determination of what makes bird genomes smaller than those of other vertebrates (12); and the genetics of some specialized traits, such as flight, tooth loss (14), blood and platelet cell genes (15, 16), sex chromosome determination (17), and vocal learning (12, 18–20). The latter trait is convergent in songbirds, parrots, and hummingbirds and is the main behavioral basis of human spoken language (21).

Partway through the project we had difficulty generating high-resolution branches on the avian family tree using protein coding genes. Experts in phylogeny indicated that although significant effort and expense went toward sequencing 45 new avian genomes, the number of species representing deep divergent nodes in the tree still may not be enough to obtain high resolution on all branches, i.e., not enough taxa sampling. In discussions in 2011 with the BGI CEO and Director, Wang Jun, about the possible taxa sampling issue, a decision was made to start an even more ambitious project, which we called the B10K project, the mission of which was to sequence representative genomes of all ~10,500 bird species (22). Later we were able to generate a higher-resolution avian family tree at the ordinal level using mostly noncoding genomic sequences, but we

Table 1 Qualitative assessment of the assembly quality and new algorithms needed to address specific tasks and questions with all genomes of a class of species with >10,000 species[a]

Task or question	High assembly quality	New algorithms > 10K species needed
Fully assemble and annotate all genes	Necessary	Yes
Resolve extant family tree	Not necessary	Yes
Redefine species	Not necessary	Yes
Reclassify species	Not necessary	Yes
Genetic elements for adaptive and specialized traits	Necessary	No
Infer history of each base in genomes	Necessary	No
Infer evolution of genome structure	Necessary	Maybe
Gaps due to extinctions	Not possible	Yes
Reconstruct common ancestor genome	Necessary	Maybe
Rescue species from extinction	Necessary	Yes
Characterize extant biodiversity	Not necessary	Yes
Predict future speciation events	Necessary	Yes

[a]The tasks/questions, the different levels of assembly quality, and the different types of algorithms needed are described in more detail in the main text. Although the accomplishment and quality of all tasks will benefit from having high-quality genome assemblies (N50 >10 Mb; with few gaps), "necessary" indicates whether high-quality assemblies are absolutely needed to complete the task. "High quality" means that nucleotide sequences of nearly all transcribed genes are completely and correctly assembled, and that chromosome organization is nearly complete and correctly assembled. "Not necessary" means that assembled genes can be incomplete and the genome structure is not needed to address the task. "Not possible" means that it is theoretically not possible to construct a high-quality genome assembly from fossilized tissue or not-well-preserved museum specimens of extinct species. "Yes" for new algorithms required means that I am not aware of existing algorithms that can handle these tasks or that new mathematical models need to be generated. "Maybe" means that I am aware of existing algorithms (as described in the main text) but am not certain if they will scale up to 10,000 or more species. "No" means I believe that there are existing algorithms that can handle these tasks at 10,000 or more species. These are qualitative assessments based on the author's experience with the Avian Phylogenomics Project and other projects.

realized that more taxa may still be needed (13). Creators of another recently inferred tree using mostly highly conserved DNA sequences claim that we had some highly supported relationships that were errors from insufficient taxa sampling (23), as they sampled 198 species and obtained some highly supported differences. However, I noticed that most of their highly conserved sequences are protein coding sequences, and some branches in their tree that are different from our genome-scale tree are more similar to our homogeneous protein coding tree that reflects exon sequence convergence (13). Even without the issue of taxa sampling, we discovered the power of having many representative genomes to address various questions in biology (11). In preparation for large-scale genomic efforts for vertebrates, and even for other classes of species, in this article I ask, what currently unanswerable questions could one answer if all extant genomes of a vertebrate class were available (**Table 1**)? I include lessons learned and predictions for the future based on the Avian Phylogenomics Project.

SETTING MINIMUM STANDARDS FOR DNA QUALITY, ASSEMBLY, AND ANNOTATION

One of the lessons learned is that, as can be expected, the quality of the genome analyses is limited by the quality of the genomic DNA, assembly, and annotations. The lower the quality of each, the

lower the quality of the results, the more effort required for analyses, and the higher the probability of errors.

As an example, for the Avian Phylogenomics Project we had three levels of genome assembly quality: high (scaffold N50 >10 Mb), intermediate (N50 between 1–10 Mb); and low (N50 <1 Mb, and typically 40 kb to 60 kb) (12). The scaffold N50 is the genome assembly fragment size in base pairs, where more than half of the genome is in scaffolds of this size or bigger. Ideally, one would want the total number of scaffolds to equal the number of chromosomes, each complete. The differences in the avian assemblies were driven mainly by financial cost and time. The high-quality assemblies were done with very expensive Sanger long-read sequences [e.g., chicken and zebra finch (5, 6)], Roche 454 or Pacbio long-read sequences [e.g., budgerigar (24)], or assisted assemblies with optical maps [e.g., budgerigar (24) and ostrich (25)]. The intermediate-quality genome assemblies were generated with Illumina short-read-only technology, with large-insert jumping libraries (0.2–40-kb insert sizes) and high coverage sequencing (60–100X). The low-quality genomes were conducted with only several short-read libraries (0.5-kb and 5-kb insert sizes) and at lower sequencing coverage (~30X). All genome assemblies were sufficient to obtain all or parts of the majority (~70%) of protein coding genes. However, some genes with low complexity sequence or GC-rich regions were not well assembled. This led to initial false conclusions of absence of such genes in our analyses, when they are really found in the unassembled raw reads or just not sequenced (or too low coverage) owing to high GC richness. Examples include several visual opsin genes (26), several dopamine receptors (27), and MHC genes (unpublished results due to inability to assemble). Only the intermediate- and high-quality genomes were sufficient for conducting analyses on genome structure and repetitive DNA evolution (**Table 1**), factors apparently important in speciation and trait evolution (28).

Going forward, to address some of the proposed questions, reduce publishing scientific errors, and eliminate the need to repeat de novo sequencing of reference genomes, minimum standards should be set for a large-scale de novo genome sequencing project of reference genomes. I propose the following: (*a*) Samples should have sufficient vouchered information and tissue to guard against species misidentification; (*b*) the heterogametic sex should be selected, as it contains both sex chromosomes and the sex chromosomes are now more easy to assemble using comparative analyses of raw reads (17); (*c*) high–molecular weight DNA needs to be isolated, preferably with sizes greater than 40–150 Kb, to accommodate long-read or mapping technologies that are needed to get through and assemble low-complexity repeat regions; (*d*) intermediate (N50 = 3–10 Mb) to high (N50 >10 Mb) assembly qualities must be generated, which is now possible and limited only by availability of high–molecular weight DNA, money, and throughput on sequencing machines for 10,000 or more species in fewer than 5–10 years; (*e*) quality-control analyses should be conducted on the assembled genomes, not only for performing standard decontamination of microorganism DNA present in tissues and in laboratories but for possible cross-contamination of tissues and sequencing samples; and (*f*) the most uniform genome annotation possible needs to be applied across species to guard against mistaking annotation differences as species differences.

The latter standard was necessary for the Avian Phylogenomics Project, as we initially had errors owing to different annotation pipelines and annotation software versions for different genomes (e.g., chicken versus zebra finch). We had to reannotate the protein coding sequences of the zebra finch and chicken genomes (the best avian genomes annotated at the time) and then use them as reference species to annotate all of the other avian species (12). However, even this approach has its limitations, as the annotations and associated genome alignments are limited to DNA sequences present in the reference species, thereby reducing annotating novel genes and genome segments unique to other species. A possible solution to this problem is the Cactus algorithm,

a reference-free genome aligner that can perform multiway species genome alignments, correct genome assemblies, and propagate uniform annotations in genomic regions that are similar across species (29).

A challenge not yet easily solved with current technologies that will cause errors in assembly and annotation is not separating assembled genomes into haplotypes (e.g., maternal and paternal chromosomes). Nearly all current genome assemblers generate only one chromosome assembly from the raw sequence reads, which is a mosaic between haplotypes. When there is a difference in the haplotype raw reads, most assemblers use a winner-takes-all rule in terms of coverage (which can be semirandom) or break the assembled region (or never stitch them together) (30). This problem is amplified in species with polyploid genomes. Resolving diploid or polyploid genomes could be possible using longer reads, chromosome optical and other mapping approaches, or new genome assembly algorithms. For this reason, the raw reads of the sequenced genomes should be kept and made available for future use when new algorithms become available, or to further help others developing such algorithms. Also, genomic DNA of the specific genome sample should be saved for use with future improved technologies.

Overall, for a large-scale genome sequencing project across an entire class of species to be successful, a set of minimum standards from tissue procurement to genome annotation must be followed (9). Such a project may require constant updating of the assemblies, alignments, and annotations. An automated updating process would be best, and probably necessary. If not done, more scientific errors will occur, and more work, money, and time will be required.

NEED FOR A CLASS-WIDE TRAIT DATABASE

If all representative genomes of an extant vertebrate class are available, they become a powerful resource for identifying the genetics of specific traits. Species differences gained through evolutionary changes serve as a natural experiment. Genome-wide association studies across species (as opposed to within a species) can be conducted. However, such experiments are greatly benefited by and sometimes require accurate and quantitative descriptions of the traits across species. In addition, the search for the associated genetic differences with a particular trait will be greatly benefitted by the ability to rule out associations with other unrelated but partially correlated traits. To accomplish such objectives, it would be best to generate a trait database of the species sequenced.

Many trait databases have been made for specific groups of species, including birds (**http://www.birdlife.org/datazone/home**; 31). However, we are not aware of one made for all species of a given class of vertebrates, and many of the databases lack the traits that many investigators are interested in. Further, not all species have been characterized for specific traits. For example, the presence or absence of vocal learning is best determined through experimental behavioral manipulations of the species, as well as through identifying the associated brain structures (32). It is assumed that there are only three vocal learning avian lineages (21), but this has not been experimentally ruled out for all orders, families, or genera of species.

To address these limitations and take full advantage of the power of comparative genomics, a single-source-trait database of all sequenced species is necessary. The database will need to include both qualitative and quantitative data for the traits. Data from existing databases could be combined, and new information could be sought and entered. The database must be highly flexible, updatable, and relational, where multiple variables can be analyzed simultaneously. The effort and expense involved in collecting more species trait data could even be higher than that required to sequence more than 10,000 genomes. However, as long as statistically sufficient numbers of

species are characterized with and without the traits of interest, the database will be sufficient to conduct comparative analyses across species. When a particular set of genetic associations of a trait are found, they can be used to predict whether some other species not yet studied has such a trait.

RESOLVE THE EXTANT CLASS FAMILY TREE

High-quality comparative genomic analyses require an accurate family tree. In addition, a main goal many will have is to use the genomes to generate the most accurate species tree possible. Theory suggests that the more species and the more assembled genomic DNA that are available, the more accurately a species tree can be inferred (33). Experiments with real data support this conclusion (13, 34). Although we overcame many challenges to generate an ordinal-level genome-scale avian species tree (13), many challenges lie ahead for generating such a tree for an entire class of vertebrates. The tree inference algorithms that we developed or scaled up to tens of genomes (13, 34, 35) will not scale to hundreds or thousands of genomes. Either they must be retooled or new algorithms must be generated (36). Algorithms to filter out misaligned genomic sequences, which would cause errors in tree inference (13), must be tooled to find out why such regions misalign (37). For this reason, new algorithms, such as Cactus (29) and UPP (an ultralarge multisequence alignment method that uses machine learning) (38), must be developed. One solution to scaling would be to generate the tree in chunks, where groups of closely related species are used to generate local trees, which are then combined into higher-level global trees. Displaying such trees is also a challenge and would require a treeview display browser that can zoom in and out of specific parts of the trees (**http://www.onezoom.org**; @TreeOfLifeApp). It was possible to display and easily visualize a tree of 48 avian species on one 8 × 11 sheet of paper or a computer screen (13), but this will not be possible with thousands or tens of thousands of species.

One of the questions that can and should be asked with having all extant species for phylogenomic inference is whether the species tree is all bifurcating, a network, or a bifurcating tree with local networks. Most studies assume that speciation is mostly if not all bifurcating. One study from the Avian Phylogenomics Project proposed that the species radiation of Neoaves was mostly bifurcating, but with local networks at major nodes owing to incomplete lineage sorting (ILS) (39). Networks can also be created by hybridization that occurs before two subspecies completely separate. With available extant genomes of an entire class, one can test these and other hypotheses on speciation and tree organization.

Such a complete tree (bifurcating or network) can then be used to determine the timing of speciation for all species in the tree. Timing is usually determined using fossils as constraints to help date the tree. However, the species classification of fossils and setting minimum and maximum constraints on a tree based on fossils can be contentious endeavors (40–43). It is not clear whether having all genomes of a vertebrate class will change contentious views, but having the structure of the tree resolved should make it easier to resolve some theories over others. One can also find out whether speciation history of different vertebrate classes follows a similar pattern after all vertebrate classes are sequenced. For example, one hypothesis is that most, if not all, modern vertebrates speciated after a mass extinction event that wiped out many of their cousin species and nearly all dinosaurs 66 Mya following an asteroid impact on Earth (44, 45). Another is that this may have happened for some groups, such as dinosaurs and mammals, but that some others, such as birds, had mass survival (42, 46).

A possible requirement for dating and resolving a tree is to have outgroups. The number of outgroup genomes may matter. When 48 bird species and 3 reptiles (crocodile, turtle, and lizard) were used, crocodiles were inferred to be sister to turtles and not to birds (13). However, when

fewer bird genomes but twice as many reptile genomes were used, including 3 crocodilian genomes, crocodiles were inferred to be sister to birds and not to turtles (47), now the more common view. Therefore, to infer a tree of a vertebrate class, as well as to perform robust comparative genomics, one must have a sufficient sample size or balance of outgroup species. Overall, with an extant tree, whether bifurcating or a combination of bifurcations and networks, researchers can take full advantage of the genomes and perform many studies not possible today. One of them is redefining the species concept.

REDEFINE SPECIES

Species, as least sexually producing ones, have been classically defined as individuals of a population that can produce fertile offspring (48, 49). However, this definition does not always apply, and it cannot be readily tested experimentally with all combinations of organisms. Morphological differences are also often used to define species, but there are many cases were morphologically distinct populations can crossbreed, generating so-called hybrids (50). The recent discovery of admixture between humans and Neanderthals, or Denisovans (51–53), has some arguing whether they should be considered separate ethnic groups or subspecies of humans instead of entirely different species (54). Conversely, cryptic species exist with no easily detectable morphological differences (55). DNA sequence differences have also been used to distinguish between species, but there is currently no formal, experimentally tested approach for this.

With the availability of the extant genomes of all "species" of a vertebrate class, one could define and redefine species based on genomic distances (56). One could calculate how distant genomic sequence identity and genome structure must be to prevent interbreeding, or to allow interbreeding resulting in infertile offspring. One could determine if the genomic distances for generating fertile offspring correlate with autosome distances, sex chromosome differences, or a combination of both. Genomic distance definitions of species will not be binomial. That is, an intermediate divergence distance might exist where individuals from two populations have more difficulty producing fertile offspring, versus infertile offspring, versus no offspring at all. If genomic distance measures are found for each of these stages of speciation, they can then be used to redefine species relationships, from the same species to subspecies level 1 (e.g., fertile hybrids), subspecies level 2 (e.g., infertile hybrids), and distinct species.

Divergence time (in thousands to millions of years) between genomes would be a secondary factor to designate species (57), but it is unlikely to be a reliable primary factor (58). Different populations can diverge at different rates, e.g., two populations diverging slowly relative to two populations diverging more rapidly.

Once created, genomic-based metrics of species definition would be useful for categorizing new species. If a hypothesized new species is discovered based on morphology or other nongenomic features, the genome of an individual can be sequenced and metrics determined to categorize it relative to other sequenced species. This approach could also be useful for resolving debates on how many species there are. As in any field, there exist lumpers and splitters. For example, the consensus view for birds is that there are approximately 10,500 species, but the number keeps rising each year as old species or subspecies are split into more species (59); some argue that there could be more than 20,000 bird subspecies (IOC World Bird List v5.3, **http://www.worldbirdnames.org**). Genomic distance analyses could help solve this debate, which might require sequencing more individuals after approximately 10,000 species are reached. It is unlikely that one would know when all species of a class have been sequenced without having a satisfactory definition of distinct "species." A class-wide sequencing project can help generate a more quantitative view of species identification.

RECLASSIFY SPECIES

Potential changes in the phylogenetic tree of the species and a more quantitative measure of genomic species relationships will invariably lead to reclassifications in taxonomy (59). One could more quantitatively group species along a classification continuum, from subspecies, species, suborder, order, subfamily, family, genera, and higher levels. If strong network relationships that overshadow bifurcating relationships were found, a new taxonomy would need to take such relationships into account. In such cases, one possibility would be to give hybrid species classifications to some species. Different groups of scientists sometimes use different criteria to classify species into, e.g., orders or families. Sequencing all extant genomes of more than one vertebrate class will lead to the ability to generate a more uniform classification across vertebrate classes. Genomic distances could be used for such a uniform criterion.

IDENTIFY GENETIC ELEMENTS FOR ADAPTIVE AND SPECIALIZED TRAITS

Once a species tree and the classification of all species are better resolved, they can be better used for comparative genomics. This includes identifying genetic elements for adaptive and specialized traits, as well as understanding genome evolution. These were the subjects of one of the flagship papers (12) and several dozen others of the Avian Phylogenomics Project, with just 48 avian species (e.g., 14, 15, 18, 20, 26, 28, 60–62). Conduct of such experiments will be greatly accelerated with the trait database of all the species.

Similar to the types of questions conducted with genomes from over 50 species (11), with over 10,000 species one can imagine more advanced comparative genomic analyses for genetic elements associated with disease resistance, loss and gain of flight, cognition, and more complex vocal learning. Tens to hundreds of species share variations on traits; for example, over 3,000 songbirds exhibit vocal learning, with abilities that range from learning only simple stereotyped songs to learning highly advanced complex songs, including some human speech (63). One might also be able to determine the genetic evolution of a trait in such a group of species, pinpointing its possible origin.

One challenge to comparative genomic trait mapping using mostly one genome per species is that some identified genetic associations could actually be specific to the individual animal rather than to the species. Such false-positive findings will occur less often when analyzing traits that exist in many species (tens to hundreds) among a vertebrate class. However, genetic changes specific to a species can be tested and further evaluated with resequencing of more individuals of that species, as was done with the crested ibis in the Avian Phylogenomics Project (64). Resequencing and comparing to a reference genome is much cheaper than performing de novo genome assembly.

INFER EVOLUTIONARY HISTORY OF EACH BASE IN THE GENOME

Many researchers are interested in the history and function of each base in the genome of a species (65, 66). This goal for a vertebrate class becomes more plausible when the genomes of all extant species are available. Outgroups would be necessary to infer the history of the bases beyond the class. Nevertheless, within a class, one can figure out more precisely which sequences are under stronger evolutionary constraint, and thus are more functionally essential. For example, there are approximately 2–3 million constrained genomic elements within both birds and mammals, but only 0.5 million are shared between birds and mammals (12). The difference is due to lineage-specific elements. The power to detect these differences depends on the number of species sequenced.

Once they are discovered, researchers can define functional elements between species and clades of species and determine whether certain bases that have been gained or lost are correlated with specific traits.

INFER THE EVOLUTION OF GENOME STRUCTURE

It has recently become apparent that genome function, including gene regulation and genome maintenance, depends on long-range interactions within the same or across distinct chromosomes, the nuclear membrane, and other molecules in the nucleus (67). Further, species evolution is also influenced by genome rearrangements (68). In the Avian Phylogenomics Project, we were able to study genome structure, organization, and rearrangements only with the intermediate to highly assembled genomes (69) (**Table 1**). With reference-quality highly assembled genomes of an entire vertebrate class, one could infer the evolution of nearly every genomic segment in a species. Further, one could study species-specific and general mechanisms of gene expression or other cell functions that depend on long-distance chromatin interactions. Then one could determine whether specific rearrangements are associated with specific traits, gene regulatory mechanisms, or susceptibility to a disease.

GAPS OWING TO EXTINCTIONS

A limitation for all of the above investigations will be the lack of genomes of extinct species, both interrelated and ancestral to the extant species. The extinct species create gaps in the extant genome-scale family tree. When the gaps are big genomic distances (i.e., deep divergence times), they make it more difficult to infer species relationships. It may be possible to fill in some gaps from recent extinctions where museum tissue samples or recent fossilized bones (<1,000,000 years) exist by sequencing the partially degraded genomes of these species (70, 71). The assembled genome quality will not likely pass the minimum standards described above. However, partial genomes are better than none, as they can be used to help anchor the species globally within a tree and for comparative genomics. For extinct species with no suitable tissue or bone samples available, fossil morphology might help. But morphological features can be subject to a high level of convergence, as recently inferred for birds (13). Thus, this challenge must be looked into more deeply. One possible compromise would be to computationally infer much of the genome of the extinct species, as discussed below.

RECONSTRUCT THE COMMON ANCESTOR GENOME OF ALL SPECIES

As one obtains more and more genomes of descendants of an individual, it becomes more and more possible to accurately reconstruct their genome, i.e., the common ancestor. A partial nucleotide reconstruction was conducted for the common ancestor genome of birds, crocodiles, and dinosaurs, i.e., an archosaur, by using several avian and reptile genomes of the Avian Phylogenomics Project (47). This was done using the Cactus program (29), which includes a module that takes the genome alignments of living species to infer at each node in the species tree the likely bases in the genome of that node. A partial genome organization reconstruction was done, using an intermediate- to high-quality avian and one reptilian (lizard) genome assembly, to infer that the extant chicken genome organization most closely resembles the dinosaur avian ancestor (69). This was done using the Evolution Highway program (**http://eh-demo.ncsa.uiuc.edu**), which uses existing genomes to take the path (i.e., highway) of least resistance to infer the chromosomal organization at each

node in the species tree. Similar types of reconstructions have been done with fewer sequenced mammalian genomes (72, 73). With all extant species of a class sequenced, one could then use both types of programs to more accurately reconstruct back in time the genomes of key nodes of ancestral species and the common ancestor of all species in the tree. With such a reconstruction, one could theoretically infer what kinds of traits that ancestor had, what allowed it to give rise to descendants that survived mass extinctions, and what adaptive genetic differences it may have had that allowed it to flourish in a particular environment at the time. One may also be able to infer whether certain genetic associations with traits found today were inherited from a common ancestor or gained or lost independently. Reconstructing the common ancestor genome could also potentially lead to genetic engineering of a reconstructed genome in living species to recapitulate the ancestral state (74).

CONTRIBUTE TO RESCUING SPECIES FROM EXTINCTION

Having representative genomes of all extant species of a class is expected to be a big boost for species conservation. The most obvious benefit is that, if one of the extant species sequenced goes extinct in the future, which is certain to occur, there will be a representative record of its genome. Another benefit will be preventing extinction of a species altogether. One study of the Avian Phylogenomics Project used the genomes of the once near-extinct crested ibis and a close cousin species, the little egret, which was not endangered (64). Comparison of their genomes with a handful of other species that were also recently near extinct, such as the bald eagle, revealed regions of the genomes of near-extinct species that may have increased susceptibility to manmade toxins, as well as inactivated genes that decreased their diet diversity. Knowledge of the toxin susceptibility and diet-related genes of each species may be useful in determining which species are more susceptible to human interference, and then minimizing such interference for them. Another benefit is that, for endangered or near-extinct species, one can conduct controlled breeding programs to increase genetic diversity, which leads to greater species survival (64, 75). Finally, theoretically, if all extant species are sequenced now, and one or more species go extinct after the genome is sequenced, representative genomes could be used for deextinction projects when such technology becomes feasible in the future (75, 76).

DETERMINE GENETIC SUSCEPTIBILITIES TO DISEASE

Similar to individuals within a human population, some species within a population are more susceptible to certain diseases than others. For example, some avian and mammalian species are more susceptible to avian influenza, a disease that they can pass onto humans, than others (77). Through comparative genomic analyses across species, one could identify the genetic elements for disease resistance. One could then genetically engineer individuals with such resistance, or develop drugs that mimic the resistance mechanism and then apply them to that species and humans. Overall, this approach could serve as a powerful tool for using natural genetic changes among species to drive biomedical discoveries for disease prevention.

CHARACTERIZE THE EXTANT BIODIVERSITY OF THE CLASS OF SPECIES

Similar to the finding that greater genetic diversity increases species survivability, greater biodiversity of a class of species increases the survival and health of that class. For example, greater

biodiversity of plants and animals leads to greater reproductivity, which is also good for the food market and human health (78, 79). With whole genomes of an entire vertebrate class, one could not only characterize the biodiversity but make a measurement of the degree of biodiversity and associate it with the health of the class, the environment, and even the planet through climate change (79).

PREDICT FUTURE SPECIATION EVENTS

Figuring out how events occurred in the past can be predictive for the future. With all extant species of a class sequenced, one could determine which species are more susceptible not only to extinction but also to greater speciation. It might even be possible to predict a future species, given certain ecological and environmental conditions, or predict what a current species will be like thousands to millions of years from now.

DEVELOP NEW ALGORITHMS FOR GENOME-SCALE ALIGNMENTS

Most of the above research areas will require development of new algorithms (**Table 1**). As mentioned above, this was the case for the Avian Phylogenomics Project at the level of 48+ genomes. We envision that enhanced or new algorithms will also be needed for storing data, generating multiway species alignments, sorting out species hybridizations from bifurcations, generating ancestral reconstructions, and performing comparative genomics for basic traits and disease, to name a few. To be efficient, this may require that the algorithms and the big data be processed on one or several servers. Computer cloud systems may be required. For example, for the Avian Phylogenomics Project it took over 400 CPU years of computing time and shipping hard drives with data to conduct many of the analyses that led to the generation of a reliable genome-scale tree of birds at the ordinal level with just 48 species (13). Scaling up to hundreds or tens of thousands of species will not be possible with some of the current algorithm and computer hardware resources. This is because many of the phylogenomic and comparative genomic approaches are NP-hard (non-deterministic polynomial-time hard) or take a long time to transport electronically. Thus, we expect the generation of representative genomes for an entire vertebrate class to spur on development of new algorithms and computer hardware.

FOSTER STRONGER RESEARCH COLLABORATIONS

Covering an entire class of vertebrate species is invariably a global effort. This is especially the case for a project with many species across all countries and many ecological niches. For the Avian Phylogenomics Project, this was accomplished by maintaining an open-door policy among participants in the project, sharing data and results, and fostering strong research collaborations within and across countries. The collaborations included over 200 researchers from over 110 institutions in 20 countries. We think that an effort to generate genomes across an entire vertebrate class will both benefit from and foster a panglobal collaboration of individuals.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank my co-coordinators of the Avian Phylogenomics Project (Guojie Zhang and Thomas Gilbert), as well as our co-coordinators of the B10K project (Gary Graves, Carsten Rahbek, and Fumin Lei) for useful discussions on the topics presented. I thank Tandy Warnow, Siavash Mirarab, Alexander Stamatakis, Simon Ho, and Benedict Paten for discussions on novel computational approaches; Wang Jun of BGI for stimulating ideas and discussions on sequencing all genomes of a class of animals; Stephen O'Brien, Klaus-Peter Koepfli, David Haussler, Oliver Ryder, Harris Lewin, Richard Durbin, Richard Ed Green, and Beth Shapiro of the G10K group for discussions of what can be done with 10,000 vertebrate species; and Peter Houde, Ed Braun, Joel Cracraft, Jon Fjeldsa, and Alexander Suh on phylogenetics. I thank chief officers, directors, and others of Pacbio, Bionano, Illumina, Dovetail, BGI, The New York Genome Center, and Duke Genome Sequencing Cores for discussions and collaborations that have led to formation of some of the opinions I expressed in this perspectives article. I thank current and former members of my laboratory, in particular Jason Howard and Carole Parent, for useful discussions on the B10K project that helped mature some of the ideas discussed. The author is supported by funds from the Howard Hughes Medical Institute.

LITERATURE CITED

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291(5507):1304–51
3. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420(6915):563–73
4. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–62
5. Int. Chick. Genome Seq. Consort. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018):695–716
6. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, et al. 2010. The genome of a songbird. *Nature* 464(7289):757–62
7. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet.* 30(9):418–26
8. Genome 10K Community Sci. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* 100(6):659–74
9. Koepfli KP, Paten B, Genome 10K Community Sci., O'Brien SJ. 2014. The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* 3:57–111
10. Li R, Fan W, Tian G, Zhu H, He L, et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* 463(7279):311–17
11. Zhang GJ, Jarvis ED, Gilbert MTP. 2014. A flock of genomes. *Science* 346(6215):1308–9
12. Zhang GJ, Li C, Li Q, Li B, Larkin DM, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215):1311–20
13. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1320–31
14. Meredith RW, Zhang G, Gilbert MT, Jarvis ED, Springer MS. 2014. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science* 346(6215):1254390
15. Opazo JC, Hoffmann FG, Natarajan C, Witt CW, Berenbrink M, Storz JF. 2015. Gene turnover in the avian globin gene families and evolutionary changes in hemoglobin isoform expression. *Mol. Biol. Evol.* 32(4):871–87
16. Ribeiro AM, Zepeda-Mendoza ML, Bertelsen MF, Kristensen AT, Jarvis ED, et al. 2015. A refined model of the genomic basis for phenotypic variation in vertebrate hemostasis. *BMC Evol. Biol.* 15:124

17. Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, et al. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* 346(6215):1246338
18. Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, et al. 2014. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* 346(6215):1256846
19. Whitney O, Pfenning AR, Howard JT, Blatti CA, Liu F, et al. 2014. Core and region-enriched networks of behaviorally regulated genes and the singing genome. *Science* 346(6215):1256780
20. Wirthlin M, Lovell PV, Jarvis ED, Mello CV. 2014. Comparative genomics reveals molecular features unique to the songbird lineage. *BMC Genom.* 15(1):1082
21. Petkov CI, Jarvis ED. 2012. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front. Evol. Neurosci.* 4:12
22. Zhang G. 2015. Genomics: Bird sequencing project takes off. *Nature* 522(7554):34
23. Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569–73
24. Ganapathy G, Howard JT, Ward JM, Li J, Li B, et al. 2014. High-coverage sequencing and annotated assemblies of the budgerigar genome. *GigaScience* 3:11
25. Zhang J, Li C, Zhou Q, Zhang G. 2015. Improving the ostrich genome assembly using optical mapping data. *GigaScience* 4:24
26. Borges R, Khan I, Johnson WE, Gilbert MTP, Zhang G, et al. 2015. Gene loss, adaptive evolution and the co-evolution of plumage coloration genes with opsins in birds. *BMC Genom.* 16(1):751
27. Haug-Baltzell A, Jarvis ED, McCarthy FM, Lyons E. 2015. Identification of dopamine receptors across the extant avian family tree and analysis with other clades uncovers a polyploid expansion among vertebrates. *Front. Neurosci.* 9:361
28. Romanov MN, Farré M, Lithgow PE, Fowler KE, Skinner BM, et al. 2014. Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. *BMC Genom.* 15(1):1060
29. Nguyen N, Hickey G, Zerbino DR, Raney B, Earl D, et al. 2015. Building a pan-genome reference for a population. *J. Comput. Biol.* 22(5):387–401
30. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, et al. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* 2(1):10
31. Rahbek C, Graves GR. 2001. Multiscale assessment of patterns of avian species richness. *PNAS* 98(8):4534–39
32. Arriaga G, Jarvis ED. 2013. Mouse vocal communication system: Are ultrasounds learned or innate? *Brain Lang.* 124(1):96–116
33. Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425(6960):798–804
34. Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346(6215):1250463
35. Kozlov AM, Aberer AJ, Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31:2577–79
36. Vachaspati P, Warnow T. 2015. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genom.* 16(Suppl. 10):S3
37. Zepeda Mendoza ML, Nygaard S, da Fonseca RR. 2014. DivA: detection of non-homologous and very divergent regions in protein sequence alignments. *BMC Res. Notes* 7:806
38. Nguyen NP, Mirarab S, Kumar K, Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.* 16:124
39. Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLOS Biol.* 13(8):e1002224
40. Ksepka DT, Ware JL, Lamm KS. 2014. Flying rocks and flying clocks: disparity in fossil and molecular dates for birds. *Proc. Biol. Sci.* 281(1788):20140677
41. Donoghue PC, Benton MJ. 2007. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. *Trends Ecol. Evol.* 22(8):424–31
42. Mitchell KJ, Cooper A, Phillips MJ. 2015. Comment on "Whole-genome analyses resolve early branches in the tree of life of modern birds." *Science* 349(6255):1460

43. Cracraft J, Houde P, Ho SYW, Mindell DP, Fjeldså J, et al. 2015. Response to Comment on "Whole-genome analyses resolve early branches in the tree of life of modern birds." *Science* 349(6255):1460
44. Schulte P, Alegret L, Arenillas I, Arz JA, Barton PJ, et al. 2010. The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science* 327(5970):1214–18
45. Benton MJ. 2010. The origins of modern biodiversity on land. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365(1558):3667–79
46. Cooper A, Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* 275(5303):1109–13
47. Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* 346(6215):1254449
48. de Queiroz K. 2005. Ernst Mayr and the modern concept of species. *PNAS* 102:6600–7
49. Silcox MT. 2014. A pragmatic approach to the species problem from a paleontological perspective. *Evol. Anthropol.* 23(1):24–26
50. Isomura N, Iwao K, Fukami H. 2013. Possible natural hybridization of two morphologically distinct species of *Acropora* (Cnidaria, Scleractinia) in the Pacific: fertilization and larval survival rates. *PLOS ONE* 8(2):e56701
51. Green RE, Kraus J, Briggs AW, Maricic T, Stenzel U, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979):710–22
52. Hawks J. 2013. Significance of Neandertal and Denisovan genomes in human evolution. *Annu. Rev. Anthropol.* 42:433–49
53. Qin P, Stoneking M. 2015. Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* 32(10):2665–74
54. Antrosio J. 2012. Denisovans, Neandertals, archaics as human races. *Living Anthropologically*, Aug. 7
55. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, et al. 2007. Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* 22(3):148–55
56. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinform.* 14:60
57. Baker RJ, Bradley RD. 2006. Speciation in mammals and the genetic species concept. *J. Mammal.* 87(4):643–62
58. Palumbi SR. 1989. Rates of molecular evolution and the fraction of nucleotide positions free to vary. *J. Mol. Evol.* 29(2):180–87
59. Sangster G. 2014. The application of species criteria in avian taxonomy and its implications for the debate over species concepts. *Biol. Rev. Camb. Philos. Soc.* 89(1):199–214
60. Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MTP, et al. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. *Genome Biol.* 15(12):539
61. Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15(12):549
62. Suh A, Churakov G, Ramakodi MP, Platt RN II, Jurka J, et al. 2015. Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol. Evol.* 7(1):205–17
63. Catchpole CK, Slater PJB. 1995. *Bird Song: Biological Themes and Variations*. Cambridge, UK: Cambridge Univ. Press
64. Li S, Li B, Cheng C, Xiong Z, Liu Q, et al. 2014. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome Biol.* 15(12):557
65. Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* 14(12):2412–23
66. Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLOS Genet.* 2(10):e173
67. Dekker J, Misteli T. 2015. Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.* In press
68. Deakin JE, Ezaz T. 2014. Tracing the evolution of amniote chromosomes. *Chromosoma* 123:201–16
69. Romanov MN, Farré M, Lithgow PE, Fowler KE, Skinner BM, et al. 2014. Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. *BMC Genom.* 15:1060

70. Orlando L, Gilbert MT, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.* 16(7):395–408
71. Der Sarkissian C, Allentoft ME, Ávila-Arcos MC, Barnett R, Campos PF, et al. 2015. Ancient genomics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370(1660):20130387
72. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309:613–17
73. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, et al. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16:1557–65
74. Lynch VJ, Bedoya-Reina OC, Ratan A, Sulak M, Drautz-Moses DI, et al. 2015. Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the Arctic. *Cell Rep.* 12(2):217–28
75. Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA. 2015. Genetic rescue to the rescue. *Trends Ecol. Evol.* 30(1):42–49
76. Loi P, Saragusty J, Ptak G. 2014. Cloning the mammoth: A complicated task or just a dream? *Adv. Exp. Med. Biol.* 753:489–502
77. Cardona CJ, Xing Z, Sandrock CE, Davis CE. 2009. Avian influenza in birds and mammals. *Comp. Immunol. Microbiol. Infect. Dis.* 32(4):255–73
78. Carroll SP, Jørgensen PS, Kinnison MT, Bergstrom CT, Denison RF, et al. 2014. Applying evolutionary biology to address global challenges. *Science* 346(6207):1245993
79. Cheng A, Mayes S, Dalle G, Demissew S, Massawe F. 2015. Diversifying crops for food and nutrition security—a case of teff. *Biol. Rev. Camb. Philos. Soc.* In press

Annual Review of
Animal Biosciences

Volume 4, 2016

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Animal Biosciences* articles may be found at http://www.annualreviews.org/errata/animal