A Case for Evolutionary Genomics and the Comprehensive Examination of Sequence Biodiversity

David D. Pollock,† Jonathan A. Eisen,‡ Norman A. Doggett,§ and Michael P. Cummings‖*

*Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico; †Department of Biological Sciences, Louisiana State University at Baton Rouge; ‡Institute for Genomic Research, Gaithersburg, Maryland; §Genomics Group, Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico; and ‖The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts

Comparative analysis is one of the most powerful methods available for understanding the diverse and complex systems found in biology, but it is often limited by a lack of comprehensive taxonomic sampling. Despite the recent development of powerful genome technologies capable of producing sequence data in large quantities (witness the recently completed first draft of the human genome), there has been relatively little change in how evolutionary studies are conducted. The application of genomic methods to evolutionary biology is a challenge, in part because gene segments from different organisms are manipulated separately, requiring individual purification, cloning, and sequencing. We suggest that a feasible approach to collecting genome-scale data sets for evolutionary biology (i.e., evolutionary genomics) may consist of combination of DNA samples prior to cloning and sequencing, followed by computational reconstruction of the original sequences. This approach will allow the full benefit of automated protocols developed by genome projects to be realized; taxon sampling levels can easily increase to thousands for targeted genomes and genomic regions. Sequence diversity at this level will dramatically improve the quality and accuracy of phylogenetic inference, as well as the accuracy and resolution of comparative evolutionary studies. In particular, it will be possible to make accurate estimates of normal evolution in the context of constant structural and functional constraints (i.e., site-specific substitution probabilities), along with accurate estimates of changes in evolutionary patterns, including pairwise coevolution between sites, adaptive bursts, and changes in selective constraints. These estimates can then be used to understand and predict the effects of protein structure and function on sequence evolution and to predict unknown details of protein structure, function, and functional divergence. In order to demonstrate the practicality of these ideas and the potential benefit for functional genomic analysis, we describe a pilot project we are conducting to simultaneously sequence large numbers of vertebrate mitochondrial genomes.

Introduction

Following the first complete genome sequence from a free-living organism (Fleischmann et al. 1995), the complete genomes of a number of crown eukaryotes have been sequenced (i.e., *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and a draft of *Homo sapiens* [*C. elegans* Consortium 1998; Adams et al. 2000; Ball et al. 2000]). The sequencing of *D. melanogaster* was particularly notable in that it was completed via whole-genome shotgun sequencing, a method previously thought applicable only to bacterial-sized genomes (Fleischmann et al. 1995; Adams et al. 2000). The human genome is also being sequenced utilizing a total-genome shotgun approach in combination with the standard directed approach (Venter, Smith, and Hood 1996; Weber and Myers 1997; Venter et al. 1998). While the data from these genomes can provide many insights into evolution, the inherent comparative nature of evolutionary biology limits the evolutionary questions that can be addressed with only these few genomes.

The paucity of genomes available from crown eukaryotes leads to common challenges for phylogenetics, molecular evolution, and functional genomics analysis, since parameters of evolutionary models, including topology and rates, will tend to have high variances as a result. This situation is particularly unacceptable for functional genomics analysis: with the completion of sequencing of the human genome, one of the most important problems of the coming century will be to develop a complete understanding of how that sequence functions to carry out essential life processes, and a major route to that understanding will be comparative analysis of sequence biodiversity. Despite the accomplishments of genomics research and the associated development of strategies, techniques, and tools for large-scale sequencing, these advances have had very limited influence on the way molecular-based evolutionary studies are conducted. In order to reduce the variance of evolutionary model parameters and increase the accuracy of comparative analysis, we need data sets from large genomic regions with much more extensive sampling of divergent taxa than in data sets that are currently available.

To address this problem, we describe here the potential challenges of evolutionary genomics, which we define as the application of genome research strategies to comparative studies in evolution. We consider possible solutions to these challenges, outline a research design for applying evolutionary genomics to vertebrate mitochondria, and describe simulated sequence assembly experiments to verify the feasibility of our design.

A key element of our plan is to reduce costs associated with cloning and handling of materials by pooling DNA samples from different gene regions and di-

Key words: evolutionary genomics, genomic biodiversity, functional genomics, comparative genomics, molecular evolution,

Address for correspondence and reprints: David D. Pollock, Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803. E-mail: daviddpollock@yahoo.com.

Mol. Biol. Evol. 17(12):1776–1788. 2000
 ISSN: 0737-4038

Table 1
Primary Products of Large-Scale Evolutionary Genomic Sequencing

- Complete sequencing of large regions from many taxa
- More accurate phylogenetic reconstruction
- Improved accuracy of estimation for evolutionary processes at and between individual sites
- Detailed comparative analysis of evolutionary processes at different sites and gene regions
- Linkage of sequence evolution to structure and function
- Production of sufficient data to improve accuracy of functional genomics predictions
- Identification of variation for population genetics analyses
- Assessment of alignment and phylogenetic reconstruction techniques

verse species prior to cloning. Thus, the costs specific to individual samples are limited to those that occur prior to cloning, and the costs of all subsequent steps are shared by all samples. Although breaking the direct association between sequences and samples is a counterintuitive approach, these associations can be re-created using automated assembly programs in combination with preexisting sequence information which can be used as an evolutionary reference. In cases where there is ambiguous assignment of sequences to samples, limited PCR and sequencing can be performed on the original samples to resolve these uncertainties. The simulated resampling and assembly experiments indicate that this strategy will be practical and cost-effective. Individuals from the same species can also be pooled but are unlikely to be separable into unique contiguous sequences, or haplotypes; rather, they would contribute to estimates of intraspecific variation.

Our focus is thus on large-scale sequence analysis of samples from divergent taxa, which we call sequence biodiversity (or genomic biodiversity in the case of complete organelle or nuclear genomes) analysis in order to distinguish it from genomic diversity studies that focus on intraspecific variation (usually in humans). While a great deal of biodiversity is currently being explored at deep taxonomic levels with the sequencing of complete bacterial genomes, these taxa are generally too divergent to be useful in evaluating many important evolutionary processes which occur on a much shorter timescale. Even with the current array of complete genomes, it is expected that around half of the genes in the human genome will be so diverged that they will not be obviously homologous to any genes of known function. Thus, what is clearly needed is greater focus on sequence biodiversity among taxa much more closely related to humans, that is, on the near-human evolutionary environment.

Areas of Impact

The primary products of a large-scale approach to evolutionary genomics are listed in table 1. The immediate products are sequences of large regions from many genomes. While there are many areas of research that would benefit from large-scale sequencing of diverse biota, we briefly describe here the possible impact on several of the most important: phylogenetics, the study of patterns and processes of sequence evolution, functional and structural genomics, and quantitative and population genetics.

Phylogenetics

Inferences concerning the historical relationships between organisms (phylogenetics) are a prerequisite for accurate comparative analyses. The accuracy of these inferences will be greatly enhanced by the data generated through evolutionary genomics in two principal ways: greater sequence information for each taxon and higher density taxon sampling. The benefits of larger amounts of sequence data per taxon for the accurate inference of phylogenetic relationships has been clearly established in studies examining the sampling properties of DNA sequence data in phylogenetic analysis (Cummings, Otto, and Wakeley 1995, 1999; Otto, Cummings, and Wakeley 1996). These studies demonstrated that accurate estimation of phylogenetic relationships required large amounts of sequence information relative to most molecular systematic studies and that significant improvements were obtained with complete mitochondrial genome sequences. The weakness of phylogenies based on single genes (Cao et al. 1994, 1998; Meyer 1994; Zardoya and Meyer 1996) and the positive effects of increased taxon sampling have also been demonstrated for both inference of taxonomic relationships and estimation of model parameters (Hillis 1996, 1998; Graybeal 1998; Kim 1998; Poe and Swofford 1999; Pollock and Bruno 2000). It has also been shown that grouping sites into clusters with similar substitution probabilities can lead to more accurate phylogenetic reconstruction (Pollock 1998), and increased taxon sampling will allow this to be done accurately in the absence of prior knowledge (Pollock and Bruno 2000). It is therefore expected with more extensive taxonomic sampling of large genomic regions that the site-specific features of evolutionary models will become more detailed and well defined and that this will lead to improved estimates of tree topology, ancestral characters, and dates of divergence between clades.

Pattern and Processes of Sequence Evolution

The study of patterns and processes of genome and sequence evolution will benefit significantly from evolutionary genomics. Currently, it is difficult to obtain suitable data sets that cover large amounts of sequence from different genes and common divergent biota. Without such data sets, one cannot do more than estimate gross differences in overall rates between sites or between subsections of phylogenetic trees, and the sensitivity of coevolutionary analysis is limited. The large increase in homologous sequences from a broad range of taxa will lead to more accurate estimation of evolutionary dynamics at individual nucleotide and amino acid positions, which is essential to detailed understanding of the forces of molecular evolution and their relationship to structure and function. These improvements in accuracy will come from the diversity of taxa directly, from the improved accuracy in topological inference,

and from having abundant sequences from identical taxa for comparison. Recent progress has been made in analyzing evolutionary behavior at individual sites, both by limiting the parameters to the equilibrium amino acid frequencies (Bruno 1996) and by optimizing functions of underlying physicochemical properties (Koshi, Mindell, and Goldstein 1999; Yang 2000). These techniques are limited, however, by the paucity of data sets that are sufficiently large for the analysis of accurate site-specific information. Likewise, analysis of the interaction between sites, or coevolution, has recently progressed to incorporate a wider diversity of coevolutionary maximum-likelihood models with reliable statistics (Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; W. J. Bruno, personal communication). In each case, accurate inferences about the patterns and processes of sequence evolution are limited by the amount of available data. Evolutionary genomics will also provide site-specific information on nucleotide and codon usage bias, insertion and deletion processes, and selective processes such as adaptation, coevolution, and functional divergence.

Functional and Structural Genomics

The increased accuracy in the estimation of evolutionary processes will enable improved correlation of natural substitutions to function and three-dimensional structure. There has been a great deal of success with attempts to make such correlations using currently available data sets (Malcolm et al. 1990; Irwin and Wilson 1991; Goldman, Thorne, and Jones 1996; Karplus et al. 1997; O'Brien, Wienberg, and Lyons 1997; Eisen 1998; Golding and Dean 1998; Clark 1999; Cort et al. 1999; D'Onofrio et al. 1999; Frishman, Goldstein, and Pollock 2000), and improved sampling of biodiversity will certainly lead to more abundant and accurate hypotheses for testing based on improved models. For example, catalytic properties of ancient ribonucleases were studied by predicting the ancestral sequences using 15 existing artiodactyl sequences (Jermann et al. 1995). Since ancestral reconstructions of this kind are known to be highly inaccurate with so few sequences (Yang, Kumar, and Nei 1995), the association of a large increase in activity with ruminant digestion, and the timing of that increase in activity, could only be helped by more accurate predictions from a larger number of ancestral sequences. The results of natural evolutionary experiments in mutagenesis and selection are inferable only indirectly, and the quality of those inferences will improve dramatically through more detailed examination of the extant products.

Functionally conserved regions of regulatory and coding regions of genes can also be efficiently identified by comparing sequences from divergent taxa, a process called evolutionary filtering (Zurawski and Clegg 1987). As with other evolutionary analyses, the power of evolutionary filtering increases rapidly with the number of related sequences examined. In the protein structural realm, it is common practice to infer the functional importance of structural regions from the sequence conservation in that region, and internalization of conserved hydrophobic regions is one of the strongest components of successful structure prediction (Livingstone and Barton 1996; Thompson and Goldstein 1996). The functional importance of structural regions can also be inferred from sequence hypervariability, as with the antigen recognition sites of HLA, plant disease resistance genes, fertilization genes, and surface antigens on viruses such as HIV and influenza. Recent work has also linked changes in evolutionary conservation and variability to functional divergence between duplicated paralogs (Zhang and Gu 1998; Gu 1999). Thus, even simple models defining differences in evolutionary rate alone, focusing on extremes of the distribution, are extremely important.

More detailed phylogenetic analyses have correlated finer structural details (such as positions of catalytic and ligand-binding sites, secondary structure features, and subunit interaction surfaces) to differences in the rate and pattern of evolutionary substitution (Goldman, Thorne, and Jones 1996, 1998; Koshi and Goldstein 1996; Thompson and Goldstein 1996, 1997; Thorne, Goldman, and Jones 1996; Koshi, Mindell, and Goldstein 1999; Dean and Golding 2000). There is every reason to believe that as the amount of evolutionary information increases dramatically, so will the accuracy in predicting structural and functional features.

The potential importance of these predictions to functional genomics cannot be understated. When the human genome is completely annotated, it can be expected that up to half of the genes will not be homologous to any genes of known function, and that many of the genes which are homologous to known genes will be far enough diverged that their exact function will be ambiguous. Large-scale analysis of expression patterns will aid in functional prediction in an efficient manner, but expensive and time-consuming experiments will eventually need to be performed. Predictions based on evolutionary genomics will provide a means of narrowing experimental possibilities to a feasible number and will allow better prediction of the possible effects of substitutions on structure and function, even when the exact structure is unknown or the gene is only remotely related to a gene of known structure.

Quantitative and Population Genetics

The inclusion of multiple individuals of the same species can provide information on the distributions and patterns of genetic polymorphisms. The simultaneous sequencing of multiple individuals has been discussed in the context of human genome research (Weber and Myers 1997), and sample pooling strategies for mutation detection have been explored (Amos, Frazier, and Wang 2000). Quantitative trait locus mapping, marker-assisted breeding, and other research using polymorphic genetic markers (i.e., single-nucleotide polymorphisms, simple sequence repeats, microsatellites, etc.) benefit from an increase in density of polymorphic markers that come from evolutionary genomics. These additional markers can be incorporated in the design of genome scanning

Table 2
Comparison of Gene-Based and Genome-Based Strategies

	Gene-Based Approach	Our Genome-Based Approach
Sampling	One gene or gene region at a time; one taxon at a time	Multiple genes or gene regions at a time; multiple taxa can be studied simultaneously
Scale	Many small individual projects funded separately with small investments, but total investment is high	Fewer large-scale projects; large overall investment in each project
Methods	Directed cloning and sequencing with minimum sequence redundancy; uncertain accuracy	Shotgun cloning of large sequences; high throughput sequencing of random shotgun clones, resulting in high sequence redundancy and good assessment of accuracy
Labor.....	Intense management of sequencing strategy: cloning and postsequence processing; redundant efforts in multiple laboratories	Little management of sequencing strategy; purification and cloning is major labor cost per taxon; low sequence management costs due to automation
Costs	Cost per base is high: a great deal of costs in producing the segments for sequencing, multiple specific primers, hand-assembly, and resequencing of ambiguities	Cost per base is low: costs are shifted from labor to automated equipment; mostly nonspecific primers in sequencing phase; automatic assembly and resequencing of gaps
Efficiency	Low	High
Automation	Limited by cost and complexity due to tracking of individual samples, primers, and short PCR products or clones	High, with a greater proportion of the work done using robotic workstations, higher throughput sequencers, and computers; human research power focused on scientific design and analysis

for mapping studies, resulting in increased statistical power.

Improved estimates of genetic diversity resulting from evolutionary genomics will also benefit population genetics research. The pattern and distribution of nucleotide diversity reflect the major forces acting on populations: mutation, selection, drift, and migration. Heterogeneity in patterns and processes of molecular evolution across genomes can be determined only through examination of multiple sequences from different genomic regions. Past studies based on extensive intraspecific-intragenomic sampling have already led to better characterization of the correlation of polymorphism and recombination (Begun and Aquadro 1992), which motivated the theory of background selection (Charlesworth, Morgan, and Charlesworth 1993; Charlesworth 1994; Charlesworth, Charlesworth, and Morgan 1995), provided a clearer understanding of the differential mutation rates between sex chromosomes and autosomes (Filatov et al. 2000), and provided better estimates of differential mutation rates between sexes (Huttley et al. 2000). A number of statistical measures and tests of natural selection have been developed based on the pattern and distribution of nucleotide diversity within and among genomes (reviewed in Tajima 1993; Clegg 1997; Wayne and Simonsen 1998). The ability to accurately estimate population genetics parameters and the power to discriminate among competing evolutionary hypotheses are directly related to the sampling employed in population genetics studies, with larger samples leading to greater accuracy and more power (Simonsen, Churchill, and Aquadro 1995). Therefore, the marginal cost of including additional individuals in evolutionary genomics projects can be weighted against the benefits of more accurate characterization of nucleotide diversity and, hence, more accurate estimates of population genetics parameters over multiple loci.

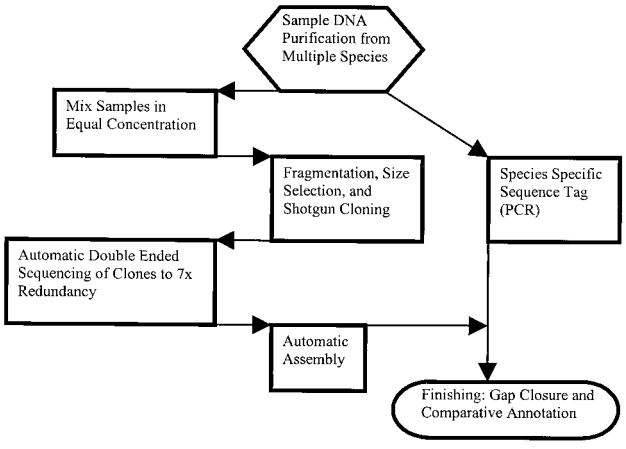

FIG. 1.—Flow chart for the proposed cloning, sequencing, and assembly procedure.

A Strategy for Moving from Small-Scale to Large-Scale Surveys of Sequence Biodiversity

The current research design model in molecular-based studies of evolution can be described as a gene-based strategy characterized by low-throughput sequencing of one or a few short regions at a time. This approach is inherently limited and does not take full advantage of developments in high-throughput technologies and associated cost efficiencies. Some principal characteristics of gene-based and genome-based strategies are compared in table 2. There are a number of procedural steps in any genomic study. These can be viewed as problems or challenges to be overcome, the solutions of which should be optimized to minimize cost without sacrificing accuracy. Analogous to arguments for whole-genome shotgun sequencing (Fleischmann et al. 1995; Venter, Smith, and Hood 1996; Weber and Myers 1997; Venter et al. 1998), we describe, in general terms, some major problems associated with moving from gene-based to genome-based research, and we discuss their solutions. We briefly describe a strategy for evolutionary genomic sequencing (fig. 1) and argue that this approach is less costly and will result in more rapid

acquisition of useful information than the gene-based strategies currently dominating evolutionary studies.

Acquisition and Preparation of Samples

In contrast to typical single-taxon genomic projects, which generally require a single sample of a common and readily available species, evolutionary genomics requires multiple samples, often from a wide diversity of taxa, some of which may not be readily available. While DNA purification using CsCl density gradient centrifugation or differential lysis may be preferable for organelle genomes when sufficient tissue is available, PCR amplification can be applied more generally and can be used when tissue samples from particular taxa are rare or expensive to collect. PCR amplification of mitochondrial genomes or large regions of nuclear DNA in a single large piece or a few smaller pieces has already been well demonstrated (Chang, Huang, and Lo 1994; Cheng et al. 1994*a*, 1994*b*; Nelson, Prodohl, and Avise 1996; Mindell et al. 1999; Miya and Nishida 1999), and these long-PCR techniques can be applied to tissues in a wide variety of preservation states. Curators and staff at several natural history museums and zoological parks (e.g., Louisiana State University, the University of California at Berkeley, the University of New Mexico, and San Diego Zoological Park) have well-documented frozen tissue collections that are sources of samples suitable for evolutionary studies.

Cloning and Sequencing Template Preparation

One factor that distinguishes a gene-based program from a genomics-based program is the size of the DNA sequence region that is examined at one time. A principal challenge is to take advantage of genomic approaches when the sequence region from any single taxon is relatively short. One way to overcome this challenge is to pool sequence regions across individuals or taxa so that the aggregate sequence corresponds to amounts well suited to genomic methods. For example, pooling 10–20 animal mitochondrial genomes, typically each about 16 kb in length, would yield an aggregate sequence comparable to a BAC, or about 0.1–0.2 the size of many bacterial genomes. Pooling ratios should be based on estimates of the number of molecules rather than DNA concentration in order to keep the concentration of each site constant in the face of length variation (e.g., differences in intron length or hypervariable regions). Pooling samples reduces the number of times labor-intensive steps involved in cloning (fractionation, size selection, ligation, transformation, etc.) have to be performed and allows for larger numbers of templates to be processed simultaneously without the need to track individuals or taxa during the sequencing phase. Pooling samples itself presents challenges with respect to sequence assembly and assignment of a specific sequence to a specific taxon or haplotype. These challenges are readily dealt with through proper taxon sampling and direct PCR of the original samples.

Sequencing

The standard procedures of genomic sequencing, including large-scale automatic sequencing of random shotgun clones, are utilized in evolutionary genomic research. Specific details such as the appropriate mixture of single- and double-ended sequencing, insert size, amount of coverage, gap closure methods, and association of contigs with taxa may vary from project to project and can be optimized to minimize cost. Our experience and limited simulation studies have indicated, however, that cost savings associated with optimizing these details are small and may be offset by costs associated with tracking clones and modifying protocols already in place at genome centers. Thus, our preferred strategy, in line with current protocols at many genomic centers, is to randomly clone inserts of 2–3 kb in length and sequence them automatically at both ends to an average redundancy of sevenfold coverage of each genomic region. This strategy minimizes costs associated with steps requiring human thought and labor, particularly the choice of clone-specific primers and final gap closure.

A critical consideration is that of identifying and minimizing the effects of sequencing errors, since they will bias estimates of variation and potentially create problems in assembly and annotation. There are several distinct sources of sequence error, depending on the source of DNA. Simply, these can be divided into error from PCR in those cases where PCR is used to generate material for cloning, and error typically associated with DNA sequencing itself. The error associated with DNA sequencing is reduced by multiple coverage of all regions sequenced. This multiple coverage for most sequence regions, sevenfold on average, is an inherent aspect of the sequencing of random shotgun clones. Furthermore, sequencing errors generally result in lower confidence in the sequencing calls (i.e., lower quality values) compared with correctly called bases, and standard assembly programs use this quality information from redundantly sequenced bases to reduce error propagation in the assembly process (Weber and Myers 1997).

PCR error due to the misincorporation of nucleotides during polymerization is expected to be similar in distribution to sequencing error. In contrast to sequencing error, however, the quality values associated with PCR errors will be indistinguishable from correct sequence, since incorrect bases will be incorporated prior to cloning. Given a large number of initial starting templates, most misincorporated bases will be of low frequency and, thus, distinguishable from the majority of authentic variation at most sites. In projects with one individual per species, there will be little natural variation (none in the case of mitochondria in the absence of heteroplasmy), and these cases can be used to obtain an accurate estimate of PCR error in the system. This estimate can then be used to correct estimates of sequence variation (particularly important in the low-frequency range) in projects with multiple individuals per species, and to calculate the probability that any given variant is

real as opposed to an artifact of PCR. For heterozygotes, both nucleotides segregating at a site should be well represented in the raw sequences and mostly distinguishable from PCR error.

Postsequencing Data Processing

Evolutionary genomics requires the same vector clipping, sequence fragment assembly, and other raw sequence data processing employed in genomics research and can be achieved using available software (i.e., Phred/Phrap/Consed, Staden Package, and/or the TIGR Assembler; Bonfield, Smith, and Staden 1995; Bonfield and Staden 1995; Sutton et al. 1995; Staden 1996; Ewing and Green 1998; Ewing et al. 1998). In projects with multiple individuals per species, mutation detection software (e.g., Bonfield, Rada, and Staden 1998) may prove useful in assigning variants to species contigs. Raw data processing that requires special applications may include sequence splitting at primers to break up chimeras resulting from ligation of PCR fragments from different genomic regions or taxa. This sequence splitting can be easily accomplished using a Perl script after vector clipping and before sequence assembly.

The task of sequence annotation is greatly simplified for evolutionary genomics projects compared to standard genomics projects because a great deal of information is available a priori with respect to the gene content and other features of the sequences being examined. This prior knowledge and availability of comparative data makes evolutionary genomics extremely well suited for annotation using case-based reasoning (Overton and Haas 1998), since there will already be a moderate number of homologous sequences available. Although appropriate software is not currently available, this prior knowledge can theoretically be used to guide sequence assembly, allowing for the ordering and linking of contigs with extremely short overlaps and thus reducing the amount of sequencing redundancy required for a minimal final gap closure effort. Subsequent steps, such as data preparation for GenBank submission, can be accomplished using available tools.

Evolutionary data analysis for inferring phylogenetic relationships, characterizing patterns of sequence evolution, and estimating population genetic parameters can also be accomplished using standard methods and available software. The only difference will be an increase in the numbers and lengths of sequences analyzed. Faster computers, improvements in algorithms (e.g., Lewis 1998), and parallel computing implementation of algorithms will continue to decrease the real time of analysis. With a rapid increase in evolutionary genomics data, there will be, however, a qualitative change in the analytical potential of the data that should spur development of novel analytical approaches.

The Case for an Examination of Vertebrate Mitochondrial Genomes

In order to give a concrete demonstration of our approach and the potential benefit for functional genomic analysis, we are implementing a pilot project to simultaneously sequence large numbers of vertebrate mitochondrial genomes. Mitochondrial genes and genomes have long been a major focus in molecular evolution, and these genomes are an excellent candidate for working out the details and demonstrating the power of evolutionary genomics. They have the advantage that they are present in high concentrations in many tissues, they are reliably amplified by PCR, and they can easily be enriched by purification of the mitochondria prior to DNA extraction (e.g., Dowling et al. 1996). The vertebrate focus of our model experiment is of rare benefit in that, unlike many primarily bacterial data sets, proteins within the vertebrate mitochondria are unlikely to have diverged to such an extent that the structural context has dramatically altered over the data set (Lesk and Chothia 1980; Chothia and Lesk 1987; Orengo et al. 1999). Thus, evolutionary analysis will not be blurred by mixing sites with widely diverged evolutionary and coevolutionary dynamics. In addition, the relatively high degree of amino acid conservation will reduce the amount of ambiguous sequence alignment, thus reducing the probability of incorrectly inferring stationary evolutionary processes for nonhomologous sites.

Mitochondrial genomes also have a strong advantage over nuclear genes in that they are unlikely to have experienced many intraspecific recombination events (but see recent controversy; Arctander 1999; Awadalla, Eyre-Walker, and Smith 1999; Merriweather and Kaestle 1999; Awadalla, Eyre-Walker, and Maynard Smith 2000; Kivisild et al. 2000). Thus, there is more likely to be a single mitochondrial phylogeny, as opposed to nuclear gene or gene segment phylogenies, which may be composites of different phylogenies. Within the vertebrates, there are few known rearrangements of genes, and none of protein-coding genes (Curole and Kocher 1999), so phylogenies are also unlikely to be confounded by inaccurate reconstruction of those events.

A detailed review concerning the effects of the current set of complete mitochondrial genomes on questions in vertebrate phylogenetics has been written (Curole and Kocher 1999), but in summary, as of March 2000, there were 69 complete vertebrate mitochondrial genomes publicly available, with slightly more than half coming from mammals (fig. 2). While many phylogenetic questions have been resolved with complete vertebrate mitochondrial genomes, many more are currently ambiguous or directly at odds with morphological and other sequence data. Thus, increasing the vertebrate mitochondrial genome data set, particularly by breaking up many of the long unbroken branches (e.g., within the rodents, bats, marsupials, snakes, lizards, turtles, and amphibia), is likely to have a large impact on confidence in the resolution of the tree structure.

The genomes of vertebrate mitochondria are also appropriate in that they contain a diversity of genes with different amounts of structural and functional information available, and with segments experiencing different structural and functional contexts and, thus, different types of selective pressure. With the cloning and sequencing of the first few mitochondrial genomes (Anderson et al. 1981, 1982*a*, 1982*b*; Bibb et al. 1981; Roe et al. 1985;

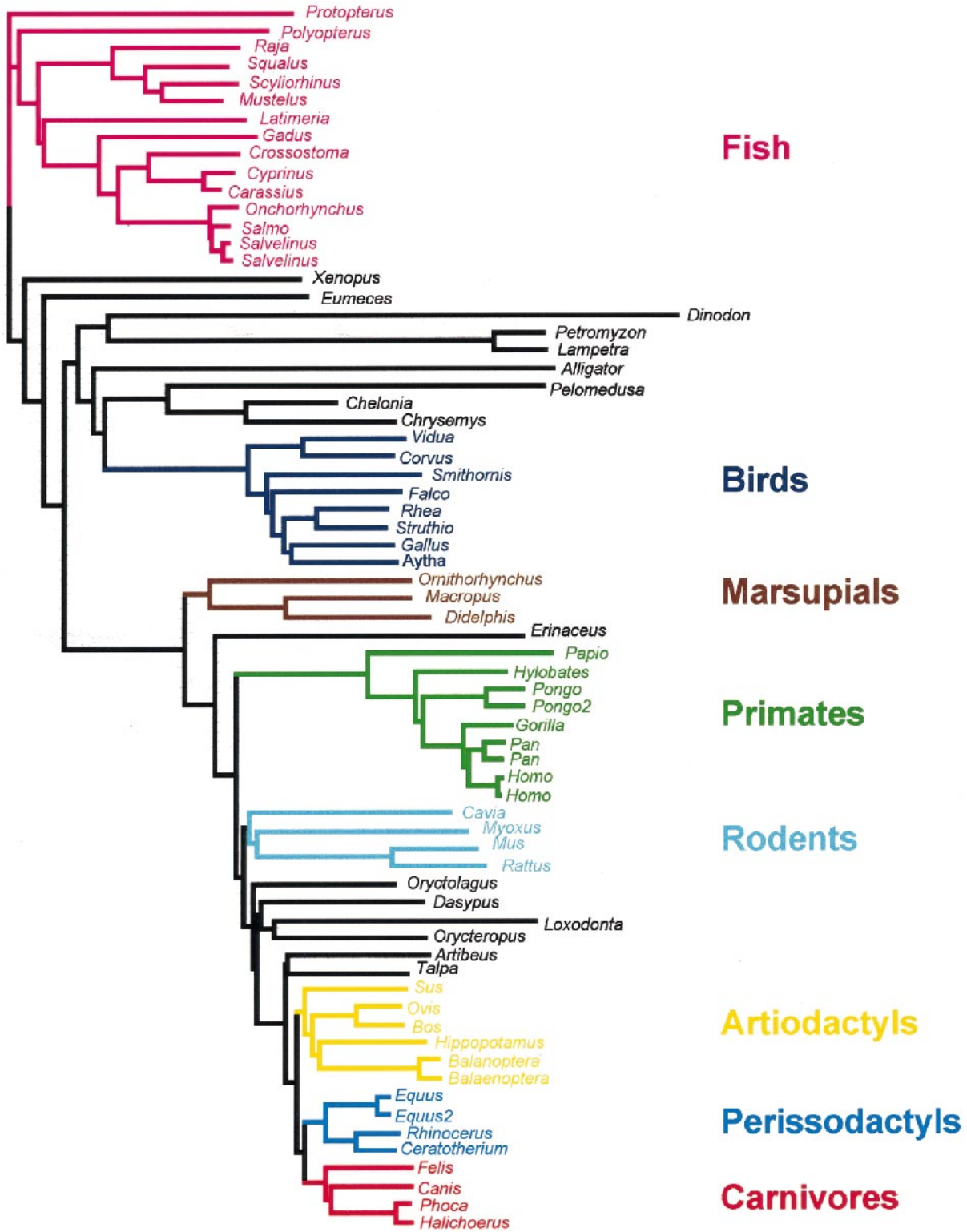

FIG. 2.—Approximate phylogenetic tree of amino acid sequences for 69 vertebrate mitochondrial genomes. Taxa are identified by their generic name, in italics, and a subset of identifiable and recognizable taxonomic clusters are labeled on the right. The tree was reconstructed using Jones-Taylor-Thornton matrix-based distances (Jones, Taylor, and Thornton 1992) and the neighbor-joining algorithm (Saitou and Nei 1987). Despite the visual rooting, the tree is an unrooted tree and is presented only as a visual approximation of the relationships among the currently available mitochondria. Bootstrap values are not given, and many of the branching relationships shown are undoubtedly incorrect.

Gadaleta et al. 1989; Desjardins and Morais 1990; Arnason, Gullberg, and Widegren 1991; Arnason and Johnsson 1992; Tzeng et al. 1992), it became clear that in addition to a control region, the vertebrate mitochondrial genome generally contains 12S and 16S ribosomal RNAs, 22 transfer RNAs, and 13 protein-coding genes that vary in length and average rate of evolution. These proteins are subunits of four different molecular complexes involved in oxidative phosphorylation and ATP synthesis: NADH reductase (NAD), cytochrome oxidase (CO), cytochrome bc_1 (CYTB), and ATP synthase (ATP). These complexes are large and have many nuclear-encoded subunits in addition to the mitochondrion-encoded proteins. Three of them, CO, CYTB, and ATP, are complete or have had many subunits crystallized, although the mitochondrion-encoded subunits of ATP (subunits a and b; ATP6 and ATP8) are the two major proteins in the complex which remain to be crystallized (Abrahams et al. 1994; Takeyasu et al. 1996; Tsukihara et al. 1996; Yoshikawa, Tsukhihara, and Shinzawa-Itoh 1996; Shirakihara et al. 1997; Uhlin, Cox, and Guss 1997; Iwata et al. 1998; Yoshikawa, Shinzawa-Itoh, and Tsukihara 1998; Vik et al. 2000). These protein subunits thus represent a wide variety of evolutionary rates, structural and functional contexts (including alpha helices, beta sheets, turns, random coils, and transmembrane regions), and examples of interaction between positions within and between protein domains and subunits. Defects in these genes have also been linked to neurological diseases, aging, and cell death, and thus there is potential to accurately relate genomic biodiversity to both normal and disease-related intraspecific genomic diversity in humans (Wallace et al. 1995).

Assessing the Feasibility of the Evolutionary Genomics Strategy: Experiments with Existing Mitochondrial Genomes

The potential assembly problems with our evolutionary genomics strategy are different from those of standard genomic projects in that misleading regions of identity in the sequence data arise not from repetitive elements, but from homologous regions in divergent taxa. The nature of the problem is very similar, however, in that these regions can lead to misassembling of contigs if they are not accounted for. In order to assess the feasibility of an evolutionary genomics strategy, computer simulations were conducted.

We tested our ability to assemble real sequences by resampling from known genomes. Sequences of 10 complete mitochondrial genomes available in GenBank were sampled following the protocol outlined above. These mitochondrial genomes were those of the human (*H. sapiens*; Anderson et al. 1981), the mouse (*Mus musculus*; Bibb et al. 1981), the cow (*Bos taurus*; Anderson et al. 1982*b*), the gorilla (*Gorilla gorilla*; Horai et al. 1995), the rat (*Rattus norvegicus*; Gadaleta et al. 1989), the finback whale (*Balaenoptera physalus*; Arnason, Gullberg, and Widegren 1991), the horse (*Equus caballus*; Xu and Arnason 1994), the domestic cat (*Felis catus*; Lopez, Cevario, and O'Brien 1996), the armadillo (*Dasypus novemcinctus*; Arnason, Gullberg, and Janke 1999), and the white rhinoceros (*Ceratotherium simum*; Xu and Arnason 1997), for a total of 183,298 bp.

To simulate the cloning process, insert size and raw sequence reads from both ends were sampled uniformly within a small range of length (2,000 kb $\pm$ 15% for insert length, 500 bp $\pm$ 5% for sequence read length) and a small range of relative concentrations of samples (plus or minus 15%). These distributions do not reflect the exact distributions of read lengths and insert sizes in genomic cloning, but they are roughly compatible with ranges observed (unpublished data); these parameters have been shown in previous simulations not to have a strong effect on simulated assembly results (Weber and Myers 1997). Standard genome projects do not have to deal with mixing of samples, so we do not have data on what the range of relative concentrations of samples will be, but the range used was intended to reasonably incorporate the variability that might be expected from DNA concentration estimates. The expected effect of a large underestimate of DNA concentration in a particular sample is that that sample will be sequenced at lower-than-optimal redundancy, and it will not be possible to fully assemble that genome. It is entirely feasible, however, that such a sample could be added at an appropriate concentration to a subsequent round for completion, while the first round could be terminated early on completion of the other genomes, thus only minimally affecting the overall cost of sequencing. Overestimates of DNA concentration in a single sample would have to be very large to dramatically affect the sequencing strategy and would generally simply lead to wasted sequence effort in direct proportion to the percentage of the overestimate. In extreme cases, sequencing would be terminated after assembly of the high-concentration sample, and the lower-concentration samples would then be recloned without that sample.

In our simulation experiment, we were able to reassemble 6 of the 10 genomes (human, cow, armadillo, whale, mouse, and horse) correctly with no gaps after sampling at 7.0-fold average coverage. There were single gaps in the rat, cat, gorilla, and rhinoceros sequences of lengths 118, 56, 63, and 90 bp, respectively, for a mean gap probability of 0.4 and a mean size of 81.5. The rhinoceros sequence was divided into two contigs, the smaller of which was 1,560 bp and bounded by the gap on one side and a 13 bp overlap on the other which was not sufficiently long to join it with the larger contig. This result is in line with expectations of this experiment for random DNA sequences, and the gaps are of a sufficient size that they could easily be closed by short PCR amplification and sequencing. Thus, the assembly was not adversely affected by the relatedness of these sequences or variation in sample concentrations, nor was it dramatically affected by substantial repetitive elements in the control regions of the horse, the rhinoceros, the cat, and the armadillo.

The taxa in this experiment were chosen to be representative of how each round in an evolutionary genomics-based vertebrate mitochondrial sequencing project might be chosen. All taxa were from separate gen-

era, but some taxon pairs, particularly the human and the gorilla, are relatively close. Are these taxa in fact representative? In order to answer this, we determined the length distribution of regions of identity for mitochondrial genomes from taxon pairs over a range of genetic distances (fig. 3). These ranged from an interclass comparison of a mammal (*H. sapiens*) and a bird (*Gallus gallus*), to the closest interspecies comparison available from the 69 taxa in figure 2, which involves *Equus asinus* (donkey) and *E. caballus*. In intragenomic comparisons within the human, gorilla, cat, and mouse genomes, identical stretches have a maximum between 14 and 17, with the exception of some repetitive expansions in the control region of the cat (*Felis catus*; data not shown). Thus, for pairwise intergenomic comparisons (fig. 3), the stretches of length 15 or more are of the greatest interest.

As expected, the most divergent comparison, that between the chicken and the human, tapered off most quickly, and there were no identical stretches longer than 35 nt (fig. 3). More closely related taxonomic pairs had more identical segments of all lengths, with the largest numbers coming from the intraequine comparison. No comparisons other than those of the intraequine (horse-donkey) and the great ape (human-gorilla) pairs had identical segments longer than 78 nt; the intraequine comparison had another 13 regions of identity of up to 93 nt, and 6 that were longer than 93 nt (lengths 97, 108, 116, 127, 128, and 205 nt), while the great ape comparison had only four regions of identity longer than 78 nt (lengths 85, 97, 127, and 164). These distributions do not theoretically present a great challenge for existing assemblers, which are designed to deal with much longer repeat sequences. Even the longest identical segment in the closest taxon pair was somewhat less than half the length of a sequence read, and less than one-tenth the length of the average cloned insert size.

Since there are roughly 4,000 species of mammals, 7,000 birds, 20,000 fish, and thousands of reptiles and amphibians, the possibilities for sampling sufficiently distinct taxa within the vertebrates are not limited; it seems feasible to avoid combining species as close as the human-gorilla pair or the horse-donkey pair, meaning that identical segments longer than 80 bp would be unlikely. Furthermore, samples from many thousands of these species are available from museums for PCR sampling, if not mitochondrial purification. A pilot project is now underway to implement the procedures we have outlined.

Evolutionary Genomics Is Cost-Effective

Any efficient method of sequencing can be used to generate the approximately 2.2 million nucleotides of raw sequence necessary in a typical experiment to simultaneously complete 20 mitochondrial genomes, and multiple laboratories worldwide could contribute to both the DNA purification and the sequencing stages. The key to cost savings is that once the cloning is done, all sequencing can be done automatically with the same two primers per clone, and there is no need to track clones. Thus, costs per individual genome are limited to sample acquisition, PCR (if applicable), and DNA purification. Assembly, filling of gaps, and verification of ambiguous organism assignment would probably be performed most efficiently at one central laboratory. Sequence can initially be assembled automatically using existing assembly programs, although this may be done more efficiently with yet-to-be-written specialty programs. A focused, large-scale evolutionary genomics effort will also avoid excess costs caused by piecemeal and sometimes redundant noncollaborative efforts in multiple laboratories.

It is our intention that assembled sequence should be made available in a dedicated phylogeny-oriented database, in addition to deposition in GenBank, in order to give added value in the form of comparative functional annotation, phylogenetic filtering, and easier access to mitochondrial-specific features. Again, a coordinated and centralized data-processing effort will avoid wasting resources through redundant efforts in multiple laboratories processing data to a point where it can be analyzed. Another large benefit of a centralized effort will be that all sequences can be tightly linked to some form of museum-deposited voucher as a matter of course, and DNA samples can also be regularly deposited with museums. Thus, one important function of a dedicated phylogenetic database will be to link directly back to the original museum information on all taxa included. This aspect of genomic analysis is often neglected by molecular biologists (few of the current complete mitochondrial genomes cite deposition of a voucher specimen), but it is extremely important to taxonomists and conservation biologists. Such linkage to original sources may not appear central to the techniques of evolutionary genomics, but it is central to creating and maintaining the resources used and to adding the full potential value of the sequence products.

Summary

We have demonstrated here that the strategy we have outlined for evolutionary genomics is theoretically viable, practical, and likely to be cost-effective. We have focused on the sampling of sequence and genomic biodiversity because the potential of these data for phylogenetics, molecular evolution, and functional genomics is compelling. At a small marginal cost increase, samples from multiple individuals in each species can also be pooled, and the intraspecific variation data obtained will then produce a large benefit for population genetics analyses. Likewise, rather than restrict the possibilities, our example of the vertebrate mitochondrial genome is intended to show the feasibility of the strategy in a specific context. The work with mitochondria is relatively straightforward, and there are fewer interpretational problems with the lack of recombination and heteroplasmy. Using the methods we have outlined here, the number of vertebrate mitochondrial genomes sequenced could easily be increased from the nearly one hundred today to thousands or more.

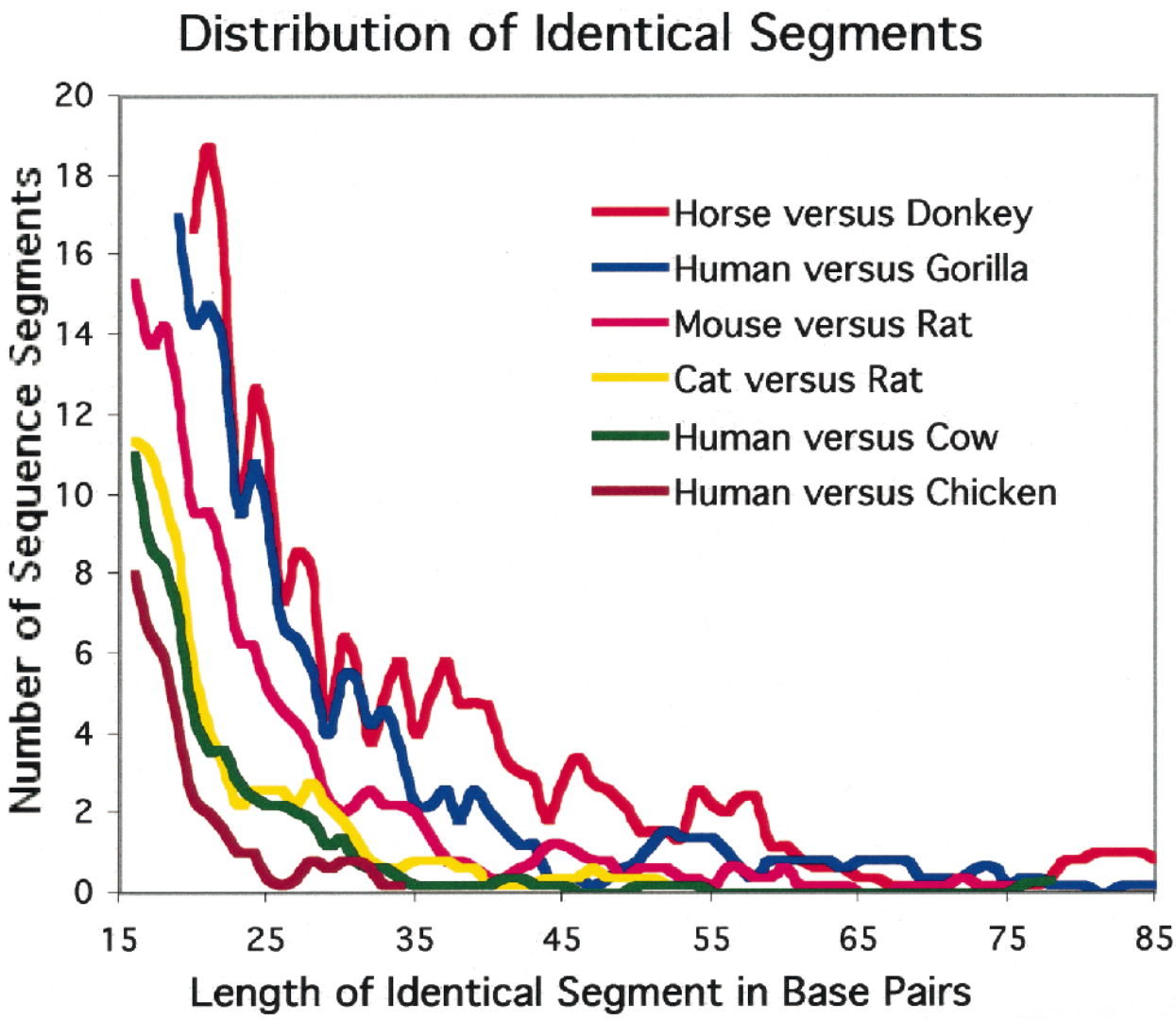

FIG. 3.—Distribution of identical segments in pairwise taxon comparisons. The number of identical segments is shown as a moving average (window size = 5 nt) for each segment length. The length of identical segments was defined as the largest possible contiguous stretch of identical nucleotides for any comparison. Comparisons were made over the entire lengths of each genome.

Given moderate funding, we believe that those thousands of taxa could be sequenced rapidly at roughly one-fifth the cost of conventional approaches. By comparison, the human genome involves about 130 times the sequencing effort with 10-fold coverage (ignoring preliminary mapping costs and duplication of efforts by noncooperating ventures). In order to put this into perspective, a single ABI 3700 automatic DNA sequencer running at full capacity could sequence 23 mitochondrial genomes per week and complete 2,000 genomes in 20 months. Thus, the full capacity of the DOE Joint Genome Institute could complete 2,000 vertebrate mitochondrial genomes in about a week, and Celera Corporation or the Human Genome Project could complete them in a matter of days. Considering this, we suggest that a moderate-sized program be initiated to sequence 2,000 or more complete genomes from vertebrate mitochondria to fully demonstrate the potential benefits of evolutionary genomics and genomic biodiversity. This would include perhaps 800, or one-fifth, of the mammals in order to focus more on the human evolutionary environment, and the remainder would be from birds, reptiles, amphibians, and fish. Ideally, such a sequencing effort would be conducted in conjunction with a dedicated collaborative bioinformatics program and would itself be viewed as a pilot for continuing community-wide large-scale nuclear evolutionary genomics and sequence biodiversity projects.

Acknowledgments

Natural history museums at Louisiana State University and the University of New Mexico contributed invaluable cooperation in making available the details of their frozen tissue collections for our analyses. C.-B. Stewart and D. Mindell both contributed expert consultation in mtDNA analysis, and we thank S. Easteal and an anonymous reviewer for helpful and insightful comments on the manuscript. D.D.P. was funded by a Director's Fellowship from Los Alamos National Labora-

tory. M.P.C. was funded by grants from NASA, NSF, and the Alfred P. Sloan Foundation.

LITERATURE CITED

ABRAHAMS, J. P., A. G. LESLIE, R. LUTTER, and J. E. WALKER. 1994. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Nature **370**:621–628.

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

AMOS, C. I., M. L. FRAZIER, and W. WANG. 2000. DNA pooling in mutation detection with reference to sequence analysis. Am. J. Hum. Genet. **66**:1689–1692.

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, and B. A. ROE. 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, and B. A. ROE. 1982*a*. Comparison of the human and bovine mitochondrial genomes. *In* P. SLONIMSKI, P. BORST, and G. ATTARDI, eds. Cold Spring Harbor monograph series. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

ANDERSON, S., M. H. L. DE BRUIJN, A. R. COULSON, I. C. EPERON, F. SANGER, and I. G. YOUNG. 1982*b*. Complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J. Mol. Biol. **156**: 683–718.

ARCTANDER, P. 1999. Mitochondrial recombination? Science **284**:2090–2091.

ARNASON, U., A. GULLBERG, and A. JANKE. 1999. The mitochondrial DNA molecule of the aardvark, *Orycteropus afer*, and the position of the Tubulidentata in the eutherian tree. Proc. R. Soc. Lond. B Biol. Sci. **266**:339–345.

ARNASON, U., A. GULLBERG, and B. WIDEGREN. 1991. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. J. Mol. Evol. **33**:556–568.

ARNASON, U., and E. JOHNSSON. 1992. The complete mitochondrial DNA sequence of the harbor seal, *Phoca vitulina*. J. Mol. Evol. **34**:493–505.

AWADALLA, P., A. EYRE-WALKER, and J. MAYNARD SMITH. 2000. Questioning evidence for recombination in human mitochondrial DNA: response. Science **288**:1931a.

AWADALLA, P., A. EYRE-WALKER, and J. M. SMITH. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286**:2524–2525.

BALL, C. A., K. DOLINSKI, S. S. DWIGHT et al. (17 co-authors). 2000. Integrating functional genomic information into the Saccharomyces Genome Database. Nucleic Acids Res. **28**: 77–80.

BEGUN, D. J., and C. F. AQUADRO. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356**:519–520.

BIBB, M. J., R. A. VAN ETTEN, C. T. WRIGHT, M. W. WALBERG, and D. A. CLAYTON. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell **26**:167–180.

BONFIELD, J. K., C. RADA, and R. STADEN. 1998. Automated detection of point mutations using fluorescent sequence trace subtraction. Nucleic Acids Res. **26**:3404–3409.

BONFIELD, J. K., K. F. SMITH, and R. STADEN. 1995. A new DNA sequence assembly program. Nucleic Acids Res. **23**: 4992–4999.

BONFIELD, J. K., and R. STADEN. 1995. The application of numerical estimates of base calling accuracy to DNA sequencing projects. Nucleic Acids Res. **23**:1406–1410.

BRUNO, W. J. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol. Biol. Evol. **13**: 1368–1374.

C. ELEGANS CONSORTIUM. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**:2012–2018.

CAO, Y., J. ADACHI, A. JANKE, S. PÄÄBO, and M. HASEGAWA. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. **39**:519–527.

CAO, Y., A. JANKE, P. J. WADDELL, M. WESTERMAN, O. TAKENAKA, S. MURATA, N. OKADA, S. PAABO, and M. HASEGAWA. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. **47**:307–322.

CHANG, Y.-S., F.-L. HUANG, and T.-B. LO. 1994. The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J. Mol. Evol. **38**:138–155.

CHARLESWORTH, B. 1994. The effect of background selection on weakly-selected, linked variants. Genet. Res. **63**:213–227.

CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289–1303.

CHARLESWORTH, D., B. CHARLESWORTH, and M. T. MORGAN. 1995. The pattern of neutral molecular variation under the background selection model. Genetics **141**:1619–1632.

CHENG, S., S.-Y. CHANG, P. GRAVITT, and R. RESPESS. 1994*a*. Long PCR. Nature **369**:684–685.

CHENG, S., C. FOCKLER, W. M. BARNES, and R. HIGUCHI. 1994*b*. Effective amplification of long targets from cloned inserts and human genomic DNA. Proc. Natl. Acad. Sci. USA **91**:5695–5699.

CHOTHIA, C., and A. M. LESK. 1987. The evolution of protein structures. Cold Spring Harb. Symp. Quant. Biol. **52**:399–406.

CLARK, M. S. 1999. Comparative genomics: the key to understanding the Human Genome Project. Bioessays **21**:121–130.

CLEGG, M. T. 1997. Plant genetic diversity and the struggle to measure selection. J. Hered. **88**:1–7.

CORT, J. R., E. V. KOONIN, P. A. BASH, and M. A. KENNEDY. 1999. A phylogenetic approach to target selection for structural genomics: solution structure of YciH. Nucleic Acids Res. **27**:4018–4027.

CUMMINGS, M. P., S. P. OTTO, and J. WAKELEY. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. **12**:814–822.

———. 1999. Genes and other samples of DNA sequence data for phylogenetic inference. Biol. Bull. **196**:345–350.

CUROLE, J. P., and T. D. KOCHER. 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. Trends Ecol. Evol. **14**:394–398.

DEAN, A. M., and G. B. GOLDING. 2000. Enzyme evolution explained (sort of). Pac. Symp. Biocomput. **5**:6–17.

DESJARDINS, P., and R. MORAIS. 1990. Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates. J. Mol. Biol. **212**:599–635.

D'ONOFRIO, G., K. JABBARI, H. MUSTO, F. ALVAREZ-VALIN, S. CRUVEILLER, and G. BERNARDI. 1999. Evolutionary genomics of vertebrates and its implications. Ann. N.Y. Acad. Sci. **870**:81–94.

DOWLING, T. E., C. MORITZ, J. D. PALMER, and L. H. RIESEBERG. 1996. Nucleic acids III: analysis of fragments and restriction sites. Sinauer, Sunderland, Mass.

EISEN, J. A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. **8**:163–167.

EWING, B., and P. GREEN. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8**:186–194.

EWING, B., L. HILLIER, M. C. WENDL, and P. GREEN. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8**:175–185.

FILATOV, D. A., F. MONEGER, I. NEGRUTIU, and D. CHARLESWORTH. 2000. Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. Nature **404**: 388–390.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (40 coauthors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science **269**:496–498, 507–512.

FRISHMAN, D., R. J. GOLDSTEIN, and D. D. POLLOCK. 2000. Protein evolution and structural genomics. Pac. Symp. Biocomput. **5**:3–5.

GADALETA, G., G. PEPE, G. DE CANDIA, C. QUAGLIARELLO, E. SBISA, and C. SACCONE. 1989. The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. J. Mol. Evol. **28**:497–516.

GOLDING, G. B., and A. M. DEAN. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol.**15**:355–369.

GOLDMAN, N., J. L. THORNE, and D. T. JONES. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. **263**: 196–208.

———. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**: 445–458.

GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol.**47**:9–17.

GU, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16**:1664–1674.

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature **383**:130–131.

———. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. **47**:3–8.

HORAI, S., K. HAYASAKA, R. KONDO, K. TSUGANE, and N. TAKAHATA. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92**:532–536.

HUTTLEY, G. A., I. B. JAKOBSEN, S. R. WILSON, and S. EASTEAL. 2000. How important is DNA replication for mutagenesis? Mol. Biol. Evol. **17**:929–937.

IRWIN, D. M., and A. C. WILSON. 1991. Structure and evolution of cow stomach lysozyme genes. FASEB J. **5**:A1527.

IWATA, S., J. W. LEE, K. OKADA, J. K. LEE, M. IWATA, B. RASMUSSEN, T. A. LINK, S. RAMASWAMY, and B. K. JAP. 1998. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. Science **281**:64–71.

JERMANN, T. M., J. G. OPITZ, J. STACKHOUSE, and S. A. BENNER. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature **374**:57–59.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

KARPLUS, K., K. SJOLANDER, C. BARRETT, M. CLINE, D. HAUSSLER, R. HUGHEY, L. HOLM, and C. SANDER. 1997. Predicting protein structure using hidden Markov models. Proteins **XX**(Suppl.):134–139.

KIM, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol.**47**:43–60.

KIVISILD, T., R. VILLEMS, L. B. JORDE, M. BAMSHAD, S. KUMAR, P. HEDRICK, T. DOWLING, M. STONEKING, T. J. PARSONS, and J. A. IRWIN. 2000. Questioning evidence for recombination in human mitochondrial DNA. Science **288**: 1931a.

KOSHI, J. M., and R. A. GOLDSTEIN. 1996. Correlating structure-dependent mutation matrices with physical-chemical properties. Pac. Symp. Biocomput. **1**:488–499.

KOSHI, J. M., D. P. MINDELL, and R. A. GOLDSTEIN. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. Mol. Biol. Evol. **16**: 173–179.

LESK, A. M., and C. CHOTHIA. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J. Mol. Biol. **136**: 225–270.

LEWIS, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol. **15**:277–283.

LIVINGSTONE, C. D., and G. J. BARTON. 1996. Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol. **266**:497–512.

LOPEZ, J. V., S. CEVARIO, and S. J. O'BRIEN. 1996. Complete nucleotide sequences of the domestic cat (*Felis catus*) mitochondrial genome and a transposed mtDNA repeat, Numt, in the nuclear genome. Genomics **33**:229–246.

MALCOLM, B. A., K. P. WILSON, B. W. MATTHEWS, J. F. KIRSCH, and A. C. WILSON. 1990. Ancestral lysozymes reconstructed neutrality tested and thermostability linked to hydrocarbon packing. Nature **345**:86–89.

MERRIWEATHER, D. A., and F. A. KAESTLE. 1999. Mitochondrial recombination? (continued). Science **285**:837.

MEYER, A. 1994. Shortcomings of the cytochrome b gene as a molecular marker. Trends Ecol. Evol. **9**:278–280.

MINDELL, D. P., M. D. SORENSON, D. E. DIMCHEFF, M. HASEGAWA, J. C. AST, and T. YURI. 1999. Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. Syst. Biol. **48**:138–152.

MIYA, M., and M. NISHIDA. 1999. Organization of the mitochondrial genome of a deep-sea fish, *Gonostoma gracile* (Teleostei: Stomiiformes): first example of transfer RNA gene rearrangements in bony fishes. Mar. Biotech. **1**:416–426.

NELSON, W. S., P. A. PRODOHL, and J. C. AVISE. 1996. Development and application of long-PCR for the assay of full-length animal mitochondrial DNA. Mol. Ecol. **5**:807–810.

O'BRIEN, S. J., J. WIENBERG, and L. A. LYONS. 1997. Comparative genomics: lessons from cats. Trends Genet. **13**: 393–399.

ORENGO, C. A., F. M. PEARL, J. E. BRAY, A. E. TODD, A. C. MARTIN, L. LE CONTE, and J. M. THORNTON. 1999. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res. **27**:275–279.

OTTO, S. P., M. P. CUMMINGS, and J. WAKELEY. 1996. Inferring phylogenies from DNA sequence data: the effects of sampling. Pp. 103–115 *in* P. H. HARVEY, A. J. LEIGH-BROWN, J. MAYNARD-SMITH, and S. NEE, eds. New uses for new phylogenies. Oxford University Press, Oxford, England.

OVERTON, G. C., and J. HAAS. 1998. Case-based reasoning driven gene annotation. Pp. 65–86 *in* S. L. SALZBERG, D. B. SEARLS, and S. KASIF, eds. Computational methods in molecular biology. Elsevier, Amsterdam.

POE, S., and D. L. SWOFFORD. 1999. Taxon sampling revisited. Nature **398**:299–300.

POLLOCK, D. D. 1998. Increased accuracy in analytical molecular distance estimation. Theor. Popul. Biol. **54**:78–90.

POLLOCK, D. D., and W. J. BRUNO. 2000. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. **17**:1854–1858.

POLLOCK, D. D., and W. R. TAYLOR. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. **10**:647–657.

POLLOCK, D. D., W. R. TAYLOR, and N. GOLDMAN. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**:187–198.

ROE, B. A., D. P. MA, R. K. WILSON, and J. F. H. WONG. 1985. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J. Biol. Chem. **260**:9759–9774.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SHIRAKIHARA, Y., A. G. LESLIE, J. P. ABRAHAMS, J. E. WALKER, T. VEDA, Y. SEKIMOTO, M. KAMBARA, K. SAIKA, Y. KAGAWA, and M. YOSHIDA. 1997. The crystal structure of the nucleotide-free alpha 3 beta 3 subcomplex of F1-ATPase from the thermophilic Bacillus PS3 is a symmetric trimer. Structure **5**:825–836.

SIMONSEN, K. L., G. A. CHURCHILL, and C. F. AQUADRO. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413–429.

STADEN, R. 1996. The Staden sequence analysis package. Mol. Biotech. **5**:233–241.

SUTTON, G., O. WHITE, M. ADAMS, and A. KERLAVAGE. 1995. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci. Tech. **1**:9–19.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

TAKEYASU, K., H. OMOTE, S. NETTIKADAN, F. TOKUMASU, A. IWAMOTO-KIHARA, and M. FUTAI. 1996. Molecular imaging of *Escherichia coli* F0F1-ATPase in reconstituted membranes using atomic force microscopy. FEBS Lett. **392**: 110–113.

THOMPSON, M. J., and R. A. GOLDSTEIN. 1996. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins **25**:38–47.

———. 1997. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. Protein Sci. **6**:1963–1975.

THORNE, J. L., N. GOLDMAN, and D. T. JONES. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

TSUKIHARA, T., H. AOYAMA, E. YAMASHITA, T. TOMIZAKI, H. YAMAGUCHI, K. SHINZAWA-ITOH, R. NAKASHIMA, R. YAONO, and S. YOSHIKAWA. 1996. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science **272**:1136–1144.

TZENG, C.-S., C.-F. HUI, S.-C. SHEN, and P. C. HUANG. 1992. The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. Nucleic Acids Res. **20**:4853–4858.

UHLIN, U., G. B. COX, and J. M. GUSS. 1997. Crystal structure of the epsilon subunit of the proton-translocating ATP synthase from *Escherichia coli.* Structure **5**:1219–1230.

VENTER, J. C., M. D. ADAMS, G. G. SUTTON, A. R. KERLAVAGE, H. O. SMITH, and M. HUNKAPILLER. 1998. Shotgun sequencing of the human genome. Science **280**:1540–1542.

VENTER, J. C., H. O. SMITH, and L. HOOD. 1996. A new strategy for genome sequencing. Nature **381**:364–366.

VIK, S. B., J. C. LONG, T. WADA, and D. ZHANG. 2000. A model for the structure of subunit a of the *Escherichia coli* ATP synthase and its role in proton translocation. Biochim. Biophys. Acta **1458**:457–466.

WALLACE, D. C., M. T. LOTT, M. D. BROWN, K. HUPOONEN, and A. TORRONI. 1995. Report of the committee on human mitochondrial DNA. Pp. 910–954 *in* A. J. CUTICHIA, ed. Human gene mapping 1995: a compendium. Johns Hopkins University Press, Baltimore, Md.

WAYNE, M. L., and K. L. SIMONSEN. 1998. Statistical tests of neutrality in the age of weak selection. Trends Ecol. Evol. **13**:236–240.

WEBER, J. L., and E. W. MYERS. 1997. Human whole-genome shotgun sequencing. Genome Res. **7**:401–409.

XU, X., and U. ARNASON. 1994. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. Gene **148**:357–362.

———. 1997. The complete mitochondrial DNA sequence of the white rhinoceros, *Ceratotherium simum,* and comparison with the mtDNA sequence of the Indian rhinoceros, *Rhinoceros unicornis.* Mol. Phylogenet. Evol. **7**:189–194.

YANG, Z. 2000. Relating physicochemical properties of amino acids to variable nucleotide substitution patterns among sites. Pac. Symp. Biocomput. **5**:78–89.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

YOSHIKAWA, S., K. SHINZAWA-ITOH, and T. TSUKIHARA. 1998. Crystal structure of bovine heart cytochrome c oxidase at 2.8 A resolution. J Bioenerg. Biomembr. **30**:7–14.

YOSHIKAWA, S., T. TSUKHIHARA, and K. SHINZAWA-ITOH. 1996. Crystal structure of fully oxidized cytochrome c-oxidase from the bovine heart at 2.8 A resolution. Biokhimiia **61**:1931–1940.

ZARDOYA, R., and A. MEYER. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol. Biol. Evol.**13**:933–942.

ZHANG, J., and X. GU. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics **149**:1615–1625.

ZURAWSKI, G., and M. T. CLEGG. 1987. Evolution of higher-plant chloroplast DNA-encoded genes implications for structure-function and phylogenetic studies. Annu. Rev. Plant Physiol. **38**:391–418.

SIMON EASTEAL, reviewing editor

Accepted August 15, 2000