

# Mitochondrial Evolution

Michael W. Gray,<sup>1\*</sup> Gertraud Burger,<sup>2</sup> B. Franz Lang<sup>2</sup>

The serial endosymbiosis theory is a favored model for explaining the origin of mitochondria, a defining event in the evolution of eukaryotic cells. As usually described, this theory posits that mitochondria are the direct descendants of a bacterial endosymbiont that became established at an early stage in a nucleus-containing (but amitochondriate) host cell. Gene sequence data strongly support a monophyletic origin of the mitochondrion from a eubacterial ancestor shared with a subgroup of the  $\alpha$ -Proteobacteria. However, recent studies of unicellular eukaryotes (protists), some of them little known, have provided insights that challenge the traditional serial endosymbiosis-based view of how the eukaryotic cell and its mitochondrion came to be. These data indicate that the mitochondrion arose in a common ancestor of all extant eukaryotes and raise the possibility that this organelle originated at essentially the same time as the nuclear component of the eukaryotic cell rather than in a separate, subsequent event.

The hypothesis of an endosymbiotic origin of the mitochondrion (1, 2), the beginnings of which surfaced over a century ago (3), draws much of its contemporary support from the discovery of a unique genome in this organelle, a relic of the mitochondrion's evolutionary past. Studies of mitochondrial DNA (mtDNA) and its expression have amply affirmed the eubacterial roots of this genome (4); mitochondrial gene sequences have enabled researchers to trace the evolutionary antecedents of mitochondria to a single ancestor related to the  $\alpha$  division of the Proteobacteria (5). Members of the rickettsial subdivision of the  $\alpha$ -Proteobacteria, a group of obligate intracellular parasites that includes genera such as *Rickettsia*, *Anaplasma*, and *Ehrlichia*, are considered to be among the closest known eubacterial relatives of mitochondria (6).

Within the past 2 years the complete sequences of the most bacteria-like mitochondrial genome [that of the protozoan *Reclinomonas americana* (7)] and the most mitochondria-like eubacterial genome [that of *Rickettsia prowazekii*, the causative agent of epidemic louse-borne typhus (8)] have been published. These two genome sequences mark the current boundaries of the evolutionary divide between mitochondria and their eubacterial relatives. Studies of these and other mitochondrial and eubacterial genomes have underpinned our attempts to understand the nature of the protomitochondrial genome from which contemporary mitochondrial genomes evolved. Mitochondrial genome se-

quences from protists (eukaryotes that are mostly unicellular, such as flagellated protozoa, amoebae, and algae) have revealed an unanticipated degree of shared primitive character (9), which provides compelling evidence that all extant mtDNAs trace their origin to a single ancestral protomitochondrial genome (see below).

## Mitochondrial DNA: Genetic Conservatism Versus Structural Diversity

As far as we know, mtDNA has the same fundamental role in all eukaryotes that contain it—namely, it encodes a limited number of RNAs and proteins essential for formation of a functional mitochondrion (10). In large part, mtDNA-specified proteins are components of respiratory complexes I (NADH: ubiquinone oxidoreductase, encoded by *nad* genes), II (succinate:ubiquinone oxidoreductase; *sdh*), III (ubiquinol:cytochrome c oxidoreductase; *cob*), and IV (cytochrome c oxidase; *cox*) of the electron transport chain as well as complex V [adenosine triphosphate (ATP) synthase; *atp*]. The organellar translation system by which mitochondrial mRNAs are decoded is also composed in part of mtDNA-specified components, notably small subunit (SSU) and large subunit (LSU) ribosomal RNAs (rRNAs) (always), 5S rRNA (rarely), and a full or partial complement of tRNAs (usually). In plants and protists, but not in animals and most fungi, some of the protein components of the mitochondrial ribosome are also encoded in mtDNA. Despite the essential coding contribution of mtDNA, most of the genetic information for mitochondrial biogenesis and function resides in the nuclear genome, with import into the organelle of nuclear DNA-specified proteins (11) and in some cases small RNAs, especially tRNAs (12).

Although the genetic role of mtDNA appears to be universally conserved, this genome exhibits remarkable variation in conformation and size as well as in actual gene content, arrangement, and expression (13). Like a typical bacterial genome, many mtDNAs map as circular molecules [but see (14)], although linear mtDNAs exist as well (15) (Fig. 1A). Mitochondrial genome size ranges from <6 kilobase pairs (kbp) in *Plasmodium falciparum* (the human malaria parasite) to >200 kbp in land plants. The mtDNA of *Arabidopsis thaliana*, a flowering plant, is the largest mitochondrial genome sequenced so far (16): at 366,924 bp, it is one-third the size (1,111,523 bp) of the genome of its eubacterial relative *R. prowazekii* (8), yet it encodes only 4% as many proteins (32 versus 834). This coding-size difference is accounted for by the extraordinarily high proportion (>80%) of noncoding sequence in *Arabidopsis* mtDNA compared with *Rickettsia* (24%). In contrast, most protist mitochondrial genomes typically have <10% of noncoding (largely intergenic spacer) sequence (9).

Gene content is similarly variable among completely sequenced mtDNAs (9, 17) (Fig. 1B). The miniscule, 5966-bp apicomplexan (*Plasmodium*) mtDNA encodes only three proteins in addition to SSU and LSU rRNAs and has no 5S rRNA or tRNA genes (Fig. 1B). Human mtDNA (16,569 bp) also lacks a 5S rRNA gene; however, it specifies 13 respiratory-chain proteins and a minimal set of tRNAs, sufficient to translate all codons. The 22-fold larger *Arabidopsis* mitochondrial genome encodes a 5S rRNA gene but only 2.5 times as many proteins as human mtDNA (32 versus 13); moreover, tRNAs specified by this spacious mtDNA are insufficient to decode the entire set of codons in the mitochondrial protein-coding genes it carries. In fact, two of the tRNA genes in *Arabidopsis* mtDNA have been recruited from the chloroplast genome in the course of evolution, constituting part of the complement of "promiscuous" chloroplast DNA sequences that comprise about 1% of this plant mitochondrial genome (16).

More genes have been found in the mitochondrial genome of *R. americana* [a heterotrophic flagellated protozoan formally described only in 1993 (18)] than in any other mtDNA characterized to date (7). *Reclinomonas* mtDNA carries a total of 97 genes, including all the protein-coding genes found in all other sequenced mtDNAs (Fig. 1B). Fully 18 protein genes of known function are unique to *Reclinomonas* mtDNA among published mitochondrial sequences, the most sur-

<sup>1</sup>Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada. <sup>2</sup>Département de Biochimie, Université de Montréal, Montréal, Québec H3C 3J7, Canada.

\*To whom correspondence should be addressed. E-mail: M.W.Gray@Dal.Ca

prising of which are four genes (*rpoA-D*) encoding a eubacteria-like, multicomponent RNA polymerase ( $\alpha_2\beta\beta'\sigma$  type). In other mitochondrial systems studied to date, transcription of mtDNA is carried out by a nucleus-encoded, single-subunit, bacteriophage T3/T7-like RNA polymerase (19), the evolutionary origin of which is obscure (20). Differences in the mechanism of mitochondrial gene expression, including oddities such as RNA editing (21) and *trans*-splicing (22), are typical of the structural and functional diversification that has accompanied evolution of mtDNA in different groups of eukaryotes.

### Ancestral Versus Derived Mitochondrial Genomes

Until recently, information about mitochondrial genome organization was sparse: relatively few complete mtDNA sequences were available, and the phylogenetic range encompassed by these mtDNAs was quite limited, with a strong bias toward animals, particularly vertebrates. From this highly skewed data set, it was difficult to deduce much about the ancestral mitochondrial genome (what we might call the protomitochondrial genome) and, in particular, to decide which contemporary mtDNAs most closely resemble the ancestral state. The contrast between the expansive plant and condensed animal mitochondrial genomes is especially striking in this regard; in virtually every parameter (for example, size, proportion of coding to noncoding sequence, rate of primary sequence divergence, conservation of gene order), these mtDNAs exhibit entirely opposite evolutionary trends. How, then, can one infer the ancestral state of the mitochondrial genome?

Within the past 6 years, many additional complete mtDNA sequences have been determined, with the list being augmented in particular through comprehensive mitochondrial genome sequencing programs focusing on the mtDNAs of fungi [Fungal Mitochondrial Genome Project (FMGP)]

(<http://megasun.bch.umontreal.ca/People/lang/FMGP/>) and protists [Organelle Genome Megasequencing Program (OGMP)] (<http://megasun.bch.umontreal.ca/ogmp/>). In the context of mitochondrial evolution, studies of protists (see Protist Image Database) (<http://megasun.bch.umontreal.ca/protists/>) are particularly important, because this group encompasses most of the phylogenetic diversity within eukaryotes (23). A recent review surveyed gene structure and gene content in 23 complete protist mtDNA sequences (9), a list that continues to grow (<http://megasun.bch.umontreal.ca/ogmp/projects/other/mtcomp.html>).

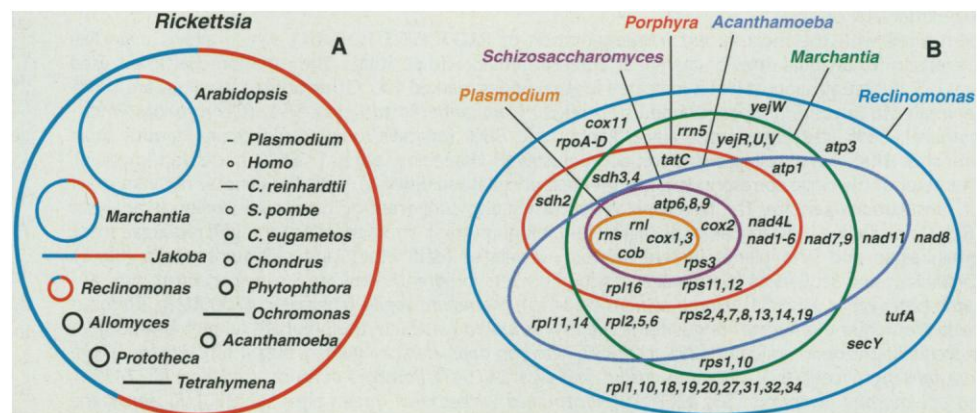
As a result of the steadily expanding database of mitochondrial genome information [see, for example, the Organelle Genome Database (GOBASE) (24)] (<http://megasun.bch.umontreal.ca/gobase/>), it is now evident that mtDNAs come in two basic types, which we designate ancestral and derived. An ancestral mitochondrial genome is defined as one that has retained clear vestiges of its eubacterial ancestry, the prototypical example being the 69,034-bp mtDNA of *R. americana* (7) (<http://megasun.bch.umontreal.ca/ogmp/projects/ramer/ramer.html>). The ancestral pattern is characterized by (i) the presence of many extra genes compared with animal mtDNA, including *sdh* and additional *nad*, *atp*, and especially ribosomal protein genes (*rps* and *rpl*) (Fig. 1B); (ii) rRNA genes that encode eubacteria-like LSU (23S), SSU (16S), and 5S rRNAs; (iii) a complete or almost complete set of tRNA genes; (iv) tight packing of genetic information in a genome that consists mostly of coding sequence, with no or few introns present; (v) eubacteria-like gene clusters; and (vi) a standard genetic code.

Derived mitochondrial genomes are ones that depart radically from the ancestral pattern, with little or no evidence of retained primitive traits, and with structural divergence usually accompanied by a substantial

reduction in overall size. Animal and most fungal mtDNAs fall into this category, as do the highly atypical mtDNAs of green algae such as *Chlamydomonas* (25) and apicomplexa such as *Plasmodium* (26). Evolution of these derived mitochondrial genomes has been marked by (i) extensive gene loss (both protein-coding and tRNA genes); (ii) marked divergence in ribosomal DNA and rRNA structure [manifested as severe truncation of rRNA sequence and secondary structure and even fragmentation of rRNA genes and dispersion of the resulting subgenomic coding modules (27)]; (iii) an accelerated rate of sequence divergence (in both protein-coding and rRNA genes); (iv) adoption of a highly biased codon usage strategy in protein genes, including in some cases wholesale elimination of certain codons (25); and (v) introduction of nonstandard codon assignments.

This is not to say that the distinction between ancestral and derived mitochondrial genomes is clear and sharp. In certain cases, departures from the prototypical ancestral pattern are evident in otherwise reasonably conserved mtDNAs. For example, in the 41,591-bp mtDNA of *Acanthamoeba castellanii* (28) (<http://megasun.bch.umontreal.ca/ogmp/projects/acast/acast.html>), a 5S rRNA gene is absent, *cox1* and *cox2* genes have become fused, many tRNA genes have been lost [with the transcripts of those that remain undergoing an unusual form of RNA editing (29)], and TGA codes for tryptophan instead of for termination. Nevertheless, the generally ancestral nature of this mitochondrial genome is evident in, for example, clustered ribosomal protein genes whose organization reflects that observed in a typical eubacterial genome (see below). Ancestral character is also evident in the greatly expanded land plant mitochondrial genomes, whose gene content approximates that of ancestral protist mtDNAs, but which have sustained a large increase in unidentified open reading frames (ORFs), noncoding sequence, and introns and

**Fig. 1.** Size and gene content of mitochondrial genomes compared with an  $\alpha$ -Proteobacterial (*Rickettsia*) genome. (A) Circles and lines represent circular and linear genome shapes, respectively. For genomes >60 kbp, the DNA coding for genes with known function (red) is distinguished from that coding for unidentified ORFs and intergenic sequences (blue). Species names are C. (*Chlamydomonas*) *reinhardtii*, C. (*Chlamydomonas*) *eugametos*, and S. (*Schizosaccharomyces*) *pombe*. (B) Gene complement of mitochondrial genomes (9). Each oval corresponds to one organism; genes included within an oval are present in the mtDNA of that organism. Only rRNA genes (*rnl*, *rns*, *rrn5*) and protein-coding genes are shown here. Full organism names, gene identities, and gene functions are listed in (9), except that *tatC* (*mttB*; also called *ymf16*) is newly recognized as a gene encoding a protein involved in membrane targeting and protein translocation (60).



intron ORFs (16).

Why and how have mitochondrial genomes evolved so differently in the various eukaryotic lineages? Currently, we are still in the initial stages of gathering the data that will ultimately provide insights into this issue. As additional mitochondrial genomes are sequenced, and the phylogenetic coverage becomes more uniform, patterns and mechanisms may begin to emerge. For example, in comparing the more ancestral mitochondrial genome of the early diverging land plant *Marchantia polymorpha* (30), with the more derived mtDNAs of recently diverged plant species such as *A. thaliana* (16) and other angiosperms, we see that the angiosperm mitochondrial genome has evolved to become recombinationally active (a condition that promotes extensive genomic rearrangements), RNA editing has appeared, and the ancestral character of the mtDNA has progressively deteriorated. This decay has involved the breakup of eubacteria-like gene clusters; fragmentation and dispersion of protein-coding genes (leading to the emergence of *trans*-splicing); gene loss (in particular ribosomal protein genes) to the nucleus; incorporation of DNA from other genomes (both chloroplast and nuclear); and progressive tRNA gene loss, compensated by import of nucleus-encoded tRNAs from the cytosol (12).

### Phylogeny of Mitochondria: Single or Multiple Origins?

In addition to information about gene content and overall genome organization, complete genome sequences yield an abundance of other data (gene order and nucleotide and amino acid sequence) that can be used to deduce evolutionary relationships. Gene order has been used to infer a mitochondrial phylogeny (31), but rampant and extensive evolutionary reshuffling of mitochondrial genes limits the usefulness of this approach. Single-gene phylogenies (especially SSU rRNA-based ones) have established many of the currently accepted affiliations among and between eubacterial, mitochondrial, and nuclear genomes (5, 6, 32, 33); however, the resolving power of single-gene analyses is limited by the inherently small information content of individual genes, complicated in the particular case of mitochondria by extreme differences in base composition and in the rate of sequence divergence of mtDNA-encoded genes in different eukaryotic lines. These rate and compositional biases result in "long branch attraction" and "base skew" artifacts, with robust (statistically supported) branching patterns contradicted by other data [see, for example, a discussion of mitochondrial SSU rRNA trees in (6)].

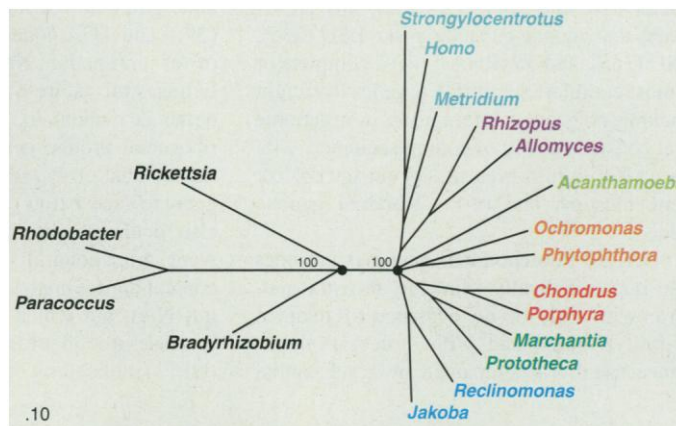
In the case of protein-coding genes, rate

difference effects are less pronounced and base bias artifacts may be substantially reduced when comparisons are carried out at the amino acid rather than at the nucleotide level. Moreover, complete genome sequencing yields sets of protein sequences that can be concatenated, providing many more informative sites than do individual proteins. An example of a mitochondrial phylogeny determined in this way, and including as outgroup all  $\alpha$ -Proteobacteria for which data are available, is shown in Fig. 2. In this analysis, in which collapsed nodes are supported by a bootstrap value of <60%, there is strong support for clades representing animals (*Homo/Strongylocentrotus/Metridium*) plus fungi (*Rhizopus/Allomyces*), green algae and land plants (*Prototheca/Marchantia*) plus red algae (*Chondrus/Porphyra*), stramenopiles (*Ochromonas/Phytophthora*), and jakobid flagellates (*Reclinomonas/Jakoba*); however, the relative order in which these mitochondrial clades and other individual mitochondrial lineages (for example, *Acanthamoeba*) diverge cannot be determined. This emerging picture inferred from mitochondrial genome data is remarkably congruent with a recent proposal (34), based on nuclear gene data, of an unresolved "big bang" radiation of the various eukaryotic lineages (see below). Notably, only those clades that are well defined by mitochondrial genome analysis (Fig. 2) are also well supported by analysis of nuclear gene data.

Phylogenetic evidence derived from both SSU rRNA (6) and protein [see (8) and Fig. 2] data support the view that all mitochondrial genomes are descended from a common, protomitochondrial ancestor (that is, that the mitochondrial genome is monophyletic, implying that mitochondria originated only once in evolution). Furthermore, the genes found in various mtDNAs to date are a subset of those encoded by the *R. americana* mitochondrial genome (9) (Fig. 1B), an observation difficult to rationalize within a polyphyletic scenario in which mitochondrial genomes in different eukaryotic lines are the end result of independent reductions of much larger eubacterial genomes. Considering that the most gene-rich mtDNA known (that of *R. americana*) encodes <2% of the protein genes found in its free-living distant cousin *Escherichia coli*, convergent evolutionary reduction to virtually the same set of genes is unlikely.

More compelling evidence of common ancestry comes from comparison of gene arrangement in mtDNA. As noted above, mitochondrial gene order has been poorly conserved, but instances of retention of what must have been the ancestral arrangement are still evident in some mtDNAs. Currently, the best example of this involves genes whose homologs in *E. coli* are contained in the

**Fig. 2.** Phylogenetic relationships among mitochondria and  $\alpha$ -Proteobacteria. A concatenated, aligned data set of amino acid sequences corresponding to respiratory chain proteins apocytochrome b (Cob) and cytochrome oxidase subunits 1 to 3 (Cox1-3) was used in the analysis. Taxa include representatives of the major eukaryotic groups and all  $\alpha$ -Proteobacteria for which data are available. Phylogenetic analysis was



performed with the most recent implementation of PROTDIST/FITCH (61), which allows a Jin/Nei correction for unequal rates of change at different amino acid positions. The variation coefficient used was 0.5. Bootstrap support (%) is indicated for two nodes marked 100. Other bootstrap values are 100% (animals, *Allomyces/Rhizopus*, *Chondrus/Porphyra*, *Marchantia/Prototheca*), 96% (*Ochromonas/Phytophthora*), 95% (red algae/green algae/plants), and 88% (animals/fungi). Solid circles denote short branches that were collapsed to reflect unresolved branching order (<60% bootstrap support). Diameter of the circle corresponds to the greatest uncollapsed distance. Scale bar denotes mean number of substitutions per site. The tree topology shown is also supported by maximum-likelihood analyses (62). Color coding indicates animals (light blue), fungi (purple), stramenopiles (orange), red algae (red), green algae and land plants (green), jakobid flagellates (dark blue), and  $\alpha$ -Proteobacteria (black). Organisms and sequences (GenBank accession numbers in parentheses) are *Strongylocentrotus purpuratus* (sea urchin; X12631), *Homo sapiens* (J01415), *Metridium senile* (cnidarian; AF000023), *Rhizopus stolonifer* [chytridiomycete fungus (63)], *Allomyces macrogynus* [chytridiomycete fungus; U41288], *A. castellanii* [rhizopod amoeba (28); U12386], *Ochromonas danica* [golden alga (37)], *Phytophthora infestans* [oomycete (63)], *Chondrus crispus* (red alga; Z47547), *Porphyra purpurea* (red alga; AF114794), *M. polymorpha* [liverwort (30); M68929], *Prototheca wickerhamii* (green alga; U02970), *R. americana* [jakobid flagellate (7); AF007261], *J. libera* [jakobid flagellate (37)], *Paracoccus denitrificans* (X05829, M17522, X05934, X05828), *R. prowazekii* (8) (AJ235270 to AJ235273), *Rhodobacter sphaeroides* (X56157, X62645, M57680, C45164 protein), and *Bradyrhizobium japonicum* (J03176, X68547).



contiguous *str*, *S10*, *spc*, and  $\alpha$  operons. Up to half the 32 *E. coli* genes in these operons (mostly encoding ribosomal proteins) are found in various protist and plant mtDNAs (9); in certain cases, the mitochondrial *rps* and *rpl* genes are clustered and organized in a way that faithfully mirrors the arrangement of the same genes in *E. coli* (7, 28, 35). What is particularly striking is that genes that are missing from the conserved *rps-rpl* cluster in one mtDNA are usually the same ones that are missing from this cluster in other mtDNAs. For example, whereas the arrangement *rpl2-rps19-rpl22-rps3-rpl16-rpl29-rps17* is found in both the *E. coli* and *R. prowazekii* genomes, *rpl22*, *rpl29*, and *rps17* are invariably absent from the corresponding mitochondrial clusters. Assuming independent events of gene deletion from the ancestral cluster and in the absence of any reason to suppose that particular mtDNA-encoded ribosomal protein genes have been preferentially lost or retained (a situation that might predispose to evolutionary convergence upon a common organizational pattern), we interpret these shared deletions as an indication that the corresponding genes had already been lost in a common mitochondrial ancestor (6, 7, 9). These particular mitochondrion-specific *rps-rpl* deletions have now been documented in a phylogenetically broad range of eukaryotes, including a land plant (*M. polymorpha*) (30), a green alga (*Nephroselmis olivacea*) (35), three amoeboid protozoa [*A. castellanii* (28), *Dictyostelium discoideum* (36), *Naegleria gruberi* (37)], and two jakobid flagellates [*R. americana* (7) and *Jakoba libera* (37)]. These observations support the view that the mitochondrial genomes of these organisms shared a common ancestor in which these deletions had already occurred.

### Endosymbiosis and Genome Reduction

As a consequence of their endosymbiotic lifestyle, parasitic eubacteria tend to have substantially reduced genomes compared with their free-living relatives (8, 38). In particular, genes for amino acid biosynthesis, nucleoside biosynthesis, anaerobic glycolysis, and regulation appear to be most at risk of deletion from the parasite's genome because such functions either become dispensable or can be complemented by host (nucleus encoded) functions (8). Comparison of the genomes of *R. prowazekii* and *R. americana* mitochondria (834 versus 62 protein-coding genes, respectively) identifies additional genes that must have been lost during evolution of the mitochondrial genome. Among the *Rickettsia* genes that are not found in mtDNA are genes defined as operational (39) (genes involved in cofactor biosynthesis, fatty acid and phospholipid metabolism, energy and intermediary metabolism, cell envelope synthesis, and cell division) as well as most informational genes (genes directing replication, transcription, and translation). For

the most part, this "missing" genetic information is now lodged in and expressed from the nuclear genome.

The relatively low gene content of mtDNA compared with even the smallest known eubacterial genomes appears to imply a relatively rapid and extensive loss or transfer of genetic information at an early stage in the evolution of the protomitochondrial genome. Differences in gene content among extant mtDNAs are best rationalized by assuming differential gene loss after divergence from the protomitochondrial genome. Indeed, there are well-documented cases in plants of relatively recent transfer of genetic information from the mitochondrial to the nuclear genome, including both respiratory chain genes (40) and ribosomal protein genes (41). As additional complete mitochondrial genome sequences have appeared, it has also become clear that elimination of genes from mtDNA is not only an ongoing evolutionary process but that certain genes have been lost on more than one occasion. Ribosomal protein genes, for instance, are entirely absent from the mitochondrial genomes of *Plasmodium* and other apicomplexa, *Chlamydomonas* and related green algae, animals, and most fungi (9) but are inferred to have been present in the mtDNAs of the immediate evolutionary ancestors of these particular lineages. To account for this phylogenetic distribution, we must assume at least three independent losses of the ribosomal protein genes that were initially present in the protomitochondrial genome. As the database of complete mitochondrial genome sequences grows, we will be able to chart more precisely the timing and extent of gene losses and transfers from mtDNA in different eukaryotic lines.

Although "striking similarities" have been noted in the functional profiles of *Rickettsia* and mitochondria (8), there is no evidence either in the form of shared genomic characters or from phylogenetic analysis (Fig. 2) that the mitochondrial genome evolved directly from an already reduced, rickettsia-like genome. Rather, mitochondria and the rickettsial group of  $\alpha$ -Proteobacteria are almost certainly the products of separate processes of reductive genome evolution (42).

### Questioning the Serial Endosymbiosis Model of Mitochondrial Origin

The serial endosymbiosis model generally assumes that the host organism had an anaerobic, heterotrophic type of metabolism "characteristic of the eukaryotic nucleocytoplasm" (2); however, whether the host was a full-fledged, nucleus-containing eukaryote or an archaeobacterium [a possibility explicit in various formulations of the serial endosymbiosis theory (2, 43)] is less clear. Implicit in this model is the assumption that the host provided the nuclear genome of the resultant combination, with subsequent mitochondrion-to-nucleus transfer of genes related to mitochon-

drial biogenesis and function.

In fact, it has become increasingly clear that the nuclear genome, instead of having descended directly from a shared common ancestor with archaeobacteria, is an evolutionary chimera that incorporates substantial contributions from both archaeobacterial and eubacterial progenitors (44). Informational genes are largely of archaeobacterial origin, whereas operational genes appear to have come primarily from eubacteria (39). Moreover, the eubacterial component of the nuclear genome appears to be considerably greater than that usually attributed to specific gene transfer from the evolving mitochondrial genome and includes genes that have nothing to do with mitochondrial biogenesis and function (45).

To accommodate these observations, various models involving fusion of eubacterial and archaeobacterial partners in the creation of the nuclear genome have been proposed (46). Such models, although invoking a major eubacterial contribution to the nuclear genome during its initial formation, do not preclude (and in fact usually assume) a subsequent endosymbiotic acquisition of mitochondria (Fig. 3, magenta arrows).

The recognition of a group of eukaryotes, collectively termed Archezoa (47), that are both lacking in mitochondria (amitochondriate) and early diverging [according to various phylogenetic analyses, but particularly nuclear SSU rRNA trees (33, 48)] promoted the idea that the organism that played host to the mitochondrial endosymbiont was a primitive eukaryote (Fig. 3, mitochondria<sup>-</sup>), possessing both a nucleus and the ability to phagocytose. That the earliest branches of the eukaryotic phylogenetic tree are apparently populated by amitochondriate organisms has been widely interpreted to mean that these particular eukaryotes diverged away from the main eukaryotic line before the advent of mitochondria—that is, that they are primitively without mitochondria.

Recent observations have begun to challenge this view. First, genes encoding typical mitochondrial proteins (for example, chaperonins), some of which trace their ancestry to the rickettsial subdivision of the  $\alpha$ -Proteobacteria, have now been found in the nuclear genomes of those amitochondriate protists that comprise the Archezoa (49). In several cases, these "mitochondrial" proteins have been observed to reside in an organelle called the hydrogenosome (which generates ATP anaerobically, producing hydrogen as the reduced end product of its energy metabolism). These findings suggest (i) that the amitochondriate eukaryotes in question once had mitochondria but subsequently lost them [although scenarios other than transfer from the premitochondrial genome might also account for the finding of  $\alpha$ -Proteobacteria-like genes in the eukaryotic nucleus (50)] and (ii) that

the hydrogenosome and the mitochondrion have a common evolutionary origin (51).

Second, the supposition that the Archaezoa are the earliest diverging eukaryotes is increasingly being called into question. The evidence supporting such a reappraisal is strongest in the case of Microsporidia; these amitochondriate eukaryotes have now been affiliated with fungi (52)—their apparent misplacement in earlier phylogenetic trees is attributable to their rapidly evolving gene sequences. The possibility has been raised more generally that the earliest branchings of the eukaryotic tree may all be suspect for similar reasons (34, 53), with both “early” and “intermediate” branchings actually collapsing to an unresolved radiation (polychotomy) (54). The emerging revisionist view of eukaryotic evolution is a scenario characterized by a massive and virtually simultaneous radiation (big bang) at the base of the eukaryotic tree, involving virtually all extant eukaryotic phyla (34).

What effect do these recent findings have on our notions of how and when the mitochondrion originated? In SSU rRNA trees, as noted previously, many protist lineages (including most of the amitochondriate ones) diverge before the radiation of the major eukaryotic assemblages (the so-called crown eukaryotes, including animals, plants, fungi, and a number of mitochondria-containing protist groups). In addition, these trees give the impression that “molecular evolutionary distances between divergent eukaryotic taxa eclipse those observed in the entire prokaryotic world” (33). If the earliest branchings in the eukaryotic tree are largely correct, and if the ancestors of these lineages actually did

have mitochondria at one time, then we are faced with the problem that the origin of mitochondria appears to be much earlier than the time of separation of the  $\alpha$ -Proteobacteria (within which mitochondria arose) from the rest of the eubacterial lineage (55). One solution to this conundrum is to suggest that frequent endosymbioses between  $\alpha$ -Proteobacteria and eukaryotes “have led to multiple endosymbiotic origins of mitochondria” (55). Although poly-phyletic scenarios of mitochondrial origin have been considered in the recent past (56), we argue here and elsewhere (6, 9, 57) that mitochondrial genomic data increasingly favor a single origin of the mitochondrial genome. If, on the other hand, “the divergence of amitochondriate protists and crown eukaryotes is radically overestimated and actually corresponds to a very short period of time” (55), then the above time conflict between the origin of mitochondria and the divergence of the  $\alpha$ -Proteobacteria is essentially resolved. We believe this solution is the most consistent with existing data.

If the divergence of the major eukaryotic lineages, including amitochondriate ones, occurred more or less simultaneously, and if there really are no eukaryotes (and never were) that primitively lack mitochondria, then the origin of mitochondria is placed very close to, if not coincident with, the origin of the eukaryotic cell itself. In fact, an intriguing proposal, the “hydrogen hypothesis” (58), can account for a chimeric origin of the eukaryotic nucleus and the origin of anaerobic (hydrogenosomal) and ultimately aerobic (mitochondrial) energy metabolism in the same event (Fig. 3, lavender arrows). This hypothesis invokes a process of metabolic syntrophy as the driv-

ing force for an association between a hydrogen-producing eubacterial symbiont (assumed to be an  $\alpha$ -Proteobacterium) and a hydrogen-requiring archaeobacterium (the host). A similar hypothesis, but involving instead a  $\delta$ -Proteobacterial symbiont, invokes the same principle of metabolic syntrophy (59).

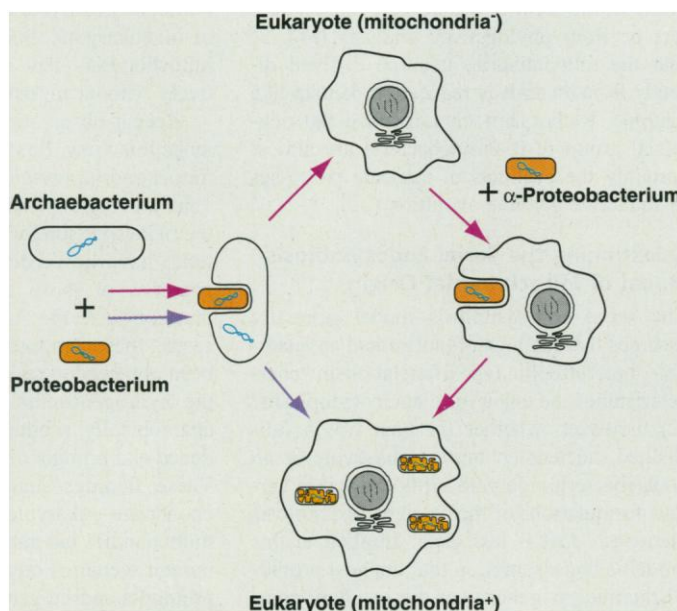
The hydrogen hypothesis (58) allows the possibility of a simultaneous origin of the ancestor of eukaryotic cells and its mitochondrion, with a major eubacterial contribution to the eukaryotic nucleus from the same  $\alpha$ -eubacterial genome whose reduction is postulated to result eventually in the mitochondrial genome (Fig. 3, lavender arrows). This hypothesis, although not excluding the possibility of a subsequent, separate endosymbiotic event leading to mitochondria (Fig. 3, magenta arrows), makes such a separate event unnecessary. The hydrogen hypothesis suggests that the origin of the mitochondrion was not only a defining event in the evolution of the eukaryotic cell but may well have been more immediately, directly, and causally related to emergence of the eukaryotic condition than is usually envisaged in the serial endosymbiosis theory.

### Concluding Remarks

Questions about mitochondrial evolution are being approached on several fronts. Systematic and phylogenetically comprehensive sequencing of mitochondrial genomes, especially from protists, has revealed much about what genes the protomitochondrial genome must have contained and how they were organized and expressed. A comparative genomics approach to mitochondrial evolution has also helped to bolster the conclusion that mitochondria are monophyletic in origin, with extant mitochondrial genomes having descended from a common protomitochondrial ancestor. The quest for mitochondrial genomes even more ancestral than that of *R. americana* continues in an effort to uncover even larger, more gene-rich mtDNAs. In fact, what this search may eventually tell us is that in those mtDNAs sequenced to date, we have already approached the upper limit of mitochondrial genetic information content. Parallel studies of eukaryotic nuclear genomes, particularly those of early-diverging protists (if these can actually be defined), may ultimately confirm whether an early, massive transfer of genetic information from an  $\alpha$ -Proteobacterial symbiont supplied much of the eubacterial complement of the nuclear genome, while producing a reduced, protomitochondrial genome from which subsequent gene loss was a much more gradual process.

The *R. prowazekii* genome sequence has solidified the connection to the mitochondrial

**Fig. 3.** Alternative hypotheses describing the origin of eukaryotic cell. Lavender arrows, simultaneous creation of the eukaryotic nucleus (gray) and mitochondrion (orange) by fusion of a hydrogen-requiring, methanogenic Archaeobacterium (host) with a hydrogen-producing  $\alpha$ -Proteobacterium (symbiont) (58). Magenta arrows, two-step scenario, initially involving formation of an amitochondriate eukaryote by fusion of an Archaeobacterium and a Proteobacterium (46) followed by acquisition of the mitochondrion through endosymbiosis with an  $\alpha$ -Proteobacterium. Bacterial and mitochondrial genomes are blue.



genome, with *Rickettsia* and mitochondrial genomes both seen as "stunning examples of highly derived genomes" (8). However, because it appears likely that reduction in genome complexity occurred independently in the rickettsial and mitochondrial lineages, albeit according to common principles, we still need to identify and study minimally diverged, free-living,  $\alpha$ -Proteobacterial relatives of mitochondria. The genome sequences of such organisms should further illuminate the reduction process underlying the transition from eubacterial to protomitochondrial genome. In this regard, given their phylogenetic position in protein trees (Fig. 2), certain rhizobial members of the  $\alpha$ -Proteobacteria (for example, *Bradyrhizobium*) are particularly interesting.

Finally, mitochondrial protein-coding sequences and genome data may ultimately help us unravel phylogenetic relationships that nuclear gene sequences are currently unable to resolve. Mitochondrial genomes comprise a cache of protein-coding genes whose origin from an  $\alpha$ -Proteobacterial ancestor is well established and whose evolution appears to track that of the eukaryotic host (there being no evidence so far of interspecies transfer of mtDNA-encoded respiratory chain or ribosomal protein genes). Determination of a wider variety of protist mtDNA sequences, together with further refinement of the eubacterial outgroup, should soon allow us to attempt the rigorous reconstruction of a eukaryotic phylogeny from mitochondrial genome data.

## References and Notes

1. L. Margulis, *Origin of Eukaryotic Cells* (Yale Univ. Press, New Haven, CT, 1970).
2. L. Margulis, *Symbiosis in Cell Evolution* (Freeman, San Francisco, 1981).
3. R. Altmann, *Die Elementarorganismen und ihre Beziehungen zu den Zellen* (Viet, Leipzig, 1890); J. Sapp, *Evolution by Association. A History of Symbiosis* (Oxford Univ. Press, New York, 1994).
4. M. W. Gray and W. F. Doolittle, *Microbiol. Rev.* **46**, 1 (1982); M. W. Gray, *Int. Rev. Cytol.* **141**, 233 (1992); N. W. Gillham, *Organelle Genes and Genomes* (Oxford Univ. Press, New York, 1994).
5. D. Yang et al., *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4443 (1985).
6. M. W. Gray and D. F. Spencer, in *Evolution of Microbial Life*, D. M. Roberts, P. Sharp, G. Alderson, M. Collins, Eds. (Cambridge Univ. Press, Cambridge, 1996), pp. 107–126.
7. B. F. Lang et al., *Nature* **387**, 493 (1997).
8. S. G. E. Andersson et al., *ibid.* **396**, 133 (1998).
9. M. W. Gray et al., *Nucleic Acids Res.* **26**, 865 (1998).
10. A. Tzagoloff and A. M. Myers, *Annu. Rev. Biochem.* **55**, 249 (1986).
11. G. Schatz and B. Dobberstein, *Science* **271**, 1519 (1996); W. Neupert, *Annu. Rev. Biochem.* **65**, 863 (1997).
12. A. Dietrich, J. H. Weil, L. Maréchal-Drouard, *Annu. Rev. Cell Biol.* **8**, 115 (1992); A. Schneider, *Trends Cell Biol.* **4**, 282 (1994).
13. D. J. Cummings, *Int. Rev. Cytol.* **141**, 1 (1992); K. Stuart and J. E. Feagin, *ibid.*, p. 65; G. D. Clark-Walker, *ibid.*, p. 89; M. R. Hanson and O. Folkerts, *ibid.*, p. 129; D. R. Wolstenholme, *ibid.*, p. 173.
14. Circular-mapping mitochondrial genomes may actually exist as multimeric linear molecules in vivo [A. J. Bendich, *Curr. Genet.* **24**, 279 (1993); D. J. Oldenburg and A. J. Bendich, *J. Mol. Biol.* **276**, 745 (1998)].
15. A. Coleman, W. F. Thompson, L. J. Goff, *J. Protozool.* **38**, 129 (1991); J. Nosek et al., *Trends Genet.* **14**, 184 (1998).
16. M. Unseld, J. R. Marienfeld, P. Brandt, A. Brennicke, *Nature Genet.* **15**, 57 (1997).
17. B. Paquin et al., *Curr. Genet.* **31**, 380 (1997).
18. M. Flavin and T. A. Nerad, *J. Eukaryotic Microbiol.* **40**, 172 (1993); C. J. O'Kelly, *ibid.*, p. 627.
19. M. W. Gray and B. F. Lang, *Trends Microbiol.* **6**, 1 (1998).
20. N. Cermakian et al., *J. Mol. Evol.* **45**, 671 (1997).
21. D. H. Price and M. W. Gray, in *Modification and Editing of RNA*, H. Grosjean and R. Benne, Eds. (ASM Press, Washington, DC, 1998), pp. 289–305; A. Marchfelder, S. Binder, A. Brennicke, V. Knoop, *ibid.*, pp. 307–323; S. L. Hajduk and R. S. Sabatini, *ibid.*, pp. 377–393; J. M. Gott and L. M. Visomirski-Robic, *ibid.*, pp. 395–411.
22. L. Bonen, *FASEB J.* **7**, 40 (1993); O. Malek and V. Knoop, *RNA* **4**, 1599 (1998).
23. D. J. Patterson and M. L. Sogin, in *The Origin and Evolution of the Cell*, H. Hartman and K. Matsuno, Eds. (World Scientific, Singapore, 1992), pp. 13–46.
24. M. Korab-Laskowska et al., *Nucleic Acids Res.* **26**, 138 (1998).
25. A. M. Nedelcu and R. W. Lee, in *The Molecular Biology of Chlamydomonas: Chloroplast and Mitochondria*, J.-D. Rochaix, Ed. (Kluwer, Dordrecht, Netherlands, 1998), pp. 63–91.
26. J. E. Feagin, *Annu. Rev. Microbiol.* **48**, 81 (1994); R. J. Wilson and D. H. Williamson, *Microbiol. Mol. Biol. Rev.* **61**, 1 (1997).
27. P. H. Boer and M. W. Gray, *Cell* **55**, 399 (1988); E. M. Denovan-Wright and R. W. Lee, *J. Biol. Chem.* **241**, 298 (1994); J. E. Feagin, B. L. Mericle, E. Werner, M. Morris, *Nucleic Acids Res.* **15**, 438 (1997).
28. G. Burger, I. Plante, K. M. Loneragan, M. W. Gray, *J. Mol. Biol.* **245**, 522 (1995).
29. K. M. Loneragan and M. W. Gray, *Science* **259**, 812 (1993); D. H. Price and M. W. Gray, *RNA* **5**, 302 (1999).
30. K. Oda et al., *J. Mol. Biol.* **223**, 1 (1992).
31. D. Sankoff et al., *Proc. Natl. Acad. Sci. U.S.A.* **89**, 6575 (1992).
32. M. W. Gray, D. Sankoff, R. J. Cedergren, *Nucleic Acids Res.* **12**, 5837 (1984); N. R. Pace, G. J. Olsen, C. R. Woese, *Cell* **45**, 325 (1986); G. J. Olsen and C. R. Woese, *FASEB J.* **7**, 113 (1993); C. R. Woese, O. Kandler, M. L. Wheelis, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576 (1990).
33. M. L. Sogin, *Curr. Opin. Genet. Dev.* **1**, 457 (1991).
34. H. Philippe and A. Adoutte, in *Evolutionary Relationships among Protozoa*, G. H. Coombs, K. Vickerman, M. A. Sleight, A. Warren, Eds. (Systematics Association, London, 1998), pp. 25–56.
35. M. W. Gray et al., in *Plant Mitochondria: From Gene to Function*, I. M. Møller, P. Gardestrom, K. Glimelius, E. Glaser, Eds. (Backhuys, Leiden, Netherlands, 1998), pp. 1–8.
36. M. Iwamoto et al., *Curr. Genet.* **33**, 304 (1998).
37. G. Burger et al., unpublished data. A summary is available at the OGMP Web site (<http://megasun.bch.umontreal.ca/ogmp/>).
38. C. M. Fraser et al., *Science* **270**, 397 (1995); R. Himmelfreich et al., *Nucleic Acids Res.* **3**, 109 (1996); C. M. Fraser et al., *Nature* **390**, 580 (1997); C. M. Fraser et al., *Science* **281**, 375 (1998); R. S. Stephens et al., *ibid.* **282**, 754 (1998).
39. M. C. Rivera, R. Jain, J. E. Moore, J. A. Lake, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6239 (1998).
40. J. M. Nugent and J. D. Palmer, *Cell* **66**, 473 (1991); P. S. Covello and M. W. Gray, *EMBO J.* **11**, 3815 (1992).
41. L. Grohmann, A. Brennicke, W. Schuster, *Nucleic Acids Res.* **20**, 5641 (1992); C. Wischmann and W. Schuster, *FEBS Lett.* **374**, 152 (1995); H. Sánchez et al., *EMBO J.* **15**, 2138 (1996); Y. Kobayashi et al., *Mol. Gen. Genet.* **256**, 589 (1997).
42. M. W. Gray, *Nature* **396**, 109 (1998); B. F. Lang, C. J. O'Kelly, G. Burger, *Protist* **149**, 313 (1998).
43. F. J. R. Taylor, *Ann. N.Y. Acad. Sci.* **503**, 1 (1987).
44. G. B. Golding and R. S. Gupta, *Mol. Biol. Evol.* **12**, 1 (1995); D.-F. Feng, G. Cho, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13028 (1997).
45. A. Markos, A. Miretsky, M. Müller, *J. Mol. Evol.* **37**, 631 (1993); P. J. Keeling and W. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1270 (1997); T. Hashimoto et al., *ibid.* **95**, 6860 (1998).
46. W. Zillig et al., *Endocytobiosis Cell Res.* **6**, 1 (1989); W. Zillig, P. Palm, H.-P. Klenk, in *The Origin and Evolution of the Cell*, H. Hartman and K. Matsuno, Eds. (World Scientific, Singapore, 1992), pp. 47–78; R. S. Gupta and G. B. Golding, *Trends Biochem. Sci.* **21**, 166 (1996).
47. T. Cavalier-Smith, in *Endocytobiology II*, W. Schwemmler and H. E. A. Schenk, Eds. (De Gruyter, Berlin, 1983), pp. 1027–1034. Members of the Archezoa include the Microsporidia (for example, *Vairimorpha necatrix*), Metamonada (diplomonads such as *Giardia lamblia*), and Parabasalida (trichomonads such as *Trichomonas vaginalis*).
48. M. L. Sogin et al., *Science* **243**, 75 (1989); C. R. Vossbrinck et al., *Nature* **362**, 411 (1989); M. L. Sogin, *Curr. Opin. Gen. Dev.* **1**, 457 (1991); D. D. Leipe, J. H. Gunderson, T. A. Nerad, M. L. Sogin, *Mol. Biochem. Parasitol.* **59**, 41 (1993); T. Cavalier-Smith, *Microbiol. Rev.* **57**, 953 (1993).
49. E. T. N. Bui, P. J. Bradley, P. J. Johnson, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 9651 (1996); A. Germot, H. Philippe, H. Le Guyader, *ibid.*, p. 14614; A. J. Roger, C. G. Clark, W. F. Doolittle, *ibid.*, p. 14618; D. S. Horner et al., *Proc. R. Soc. London Ser. B* **263**, 1053 (1996); A. Germot, H. Philippe, H. Le Guyader, *Mol. Biochem. Parasitol.* **87**, 159 (1997); R. P. Hirt et al., *Curr. Biol.* **7**, 995 (1997); A. J. Roger et al., *Proc. Natl. Acad. Sci. U.S.A.* **95**, 229 (1998); T. M. Embley and R. P. Hirt, *Curr. Opin. Genet. Dev.* **8**, 624 (1998).
50. K. Henze et al., *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9122 (1995); W. F. Doolittle, *Trends Genet.* **14**, 307 (1998).
51. M. Müller, *Parasitol. Today* **13**, 166 (1997).
52. P. J. Keeling and W. F. Doolittle, *Mol. Biol. Evol.* **13**, 1297 (1996); T. D. Edlin et al., *Mol. Phylogenet. Evol.* **5**, 359 (1996); P. J. Keeling and G. I. McFadden, *Trends Microbiol.* **6**, 19 (1998); R. P. Hirt et al., *Proc. Natl. Acad. Sci. U.S.A.* **96**, 580 (1999).
53. K. Budin and H. Philippe, *Mol. Biol. Evol.* **15**, 943 (1998); J. W. Stiller, E. C. S. Duffield, B. D. Hall, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11769 (1998); H. Philippe and J. Laurent, *Curr. Opin. Genet. Dev.* **8**, 616 (1998).
54. S. Kumar and A. Rzhetsky, *J. Mol. Evol.* **42**, 183 (1996).
55. M. L. Sogin, *Curr. Opin. Genet. Dev.* **7**, 792 (1997).
56. M. W. Gray, R. Cedergren, Y. Abel, D. Sankoff, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2267 (1989).
57. M. W. Gray, *Curr. Opin. Genet. Dev.* **3**, 884 (1993); ———, in *The Molecular Biology of Plant Mitochondria*, C. S. Leavings III and I. K. Vasil, Eds. (Kluwer, Dordrecht, Netherlands, 1995), pp. 635–659.
58. W. Martin and M. Müller, *Nature* **392**, 37 (1998).
59. D. Moreira and P. López-García, *J. Mol. Evol.* **47**, 517 (1998).
60. J. H. Weiner et al., *Cell* **93**, 93 (1998); E. G. Boggsch et al., *J. Biol. Chem.* **273**, 18003 (1998).
61. J. Felsenstein, *Phylogeny Inference Package* (version 3.6 (University of Washington, Seattle; <ftp://evolution.genetics.washington.edu/tempdir/>)).
62. J. Adachi and M. Hasegawa, *Computer Science Monograph* 28, MOLPHY version 2.3B3 (Institute of Statistical Mathematics, Tokyo, 1996); K. Strimmer and A. von Haeseler, *Mol. Biol. Evol.* **13**, 964 (1996).
63. B. F. Lang et al., unpublished data. A summary is available at the FMGP Web site (<http://megasun.bch.umontreal.ca/people/lang/FMGP/>).
64. Supported by operating grants from the Medical Research Council (MRC) of Canada (MT-4124 to M.W.G. and MT-14028 to B.F.L.) and the Canadian Genome Analysis and Technology Program (GO-12984) by a special project grant (SP-14226; OGMP) from MRC Canada and by generous equipment grants from LiCor and SUN Microsystems. M.W.G. and B.F.L. are Fellows and G.B. is an Associate in the Program in Evolutionary Biology of the Canadian Institute for Advanced Research, whom we thank for salary and interaction support.