LBNL-59357

Requirements and standards for organelle genome databases

Jeffrey L. Boore

DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, University of California Berkeley, and Genome Project Solutions

Contact information:

DOE Joint Genome Institute

2800 Mitchell Drive

Walnut Creek, CA 94598 USA

Phone: 925-296-5691

Fax: 925-296-5620

E-mail: JLBoore@berkeley.edu

Abstract

Mitochondria and plastids (collectively called organelles) descended from prokaryotes that adopted an intracellular, endosymbiotic lifestyle within early eukaryotes. Comparisons of their remnant genomes address a wide variety of biological questions, especially when including the genomes of their prokaryotic relatives and the many genes transferred to the eukaryotic nucleus during the transitions from endosymbiont to organelle. The pace of producing complete organellar genome sequences now makes it unfeasible to do broad comparisons using the primary literature and, even if it were feasible, it is now becoming uncommon for journals to accept detailed descriptions of genome-level features. Unfortunately no database is currently useful for this task, since they have little standardization and are riddled with error. Here I outline what is currently wrong and what must be done to make this data useful to the scientific community.

The several facilities built originally for the Human Genome Project have established an enormous capability for high-throughput DNA sequencing. The output of the five largest centres now tops 150 billion nucleotides per year, equivalent to 50-fold coverage of the human genome. This capacity is now being applied to the complete genome sequencing of many other organisms (see http://www.genomesonline.org/), and it appears that we will soon have in hand the genome sequences of scores of eukaryotes and many hundreds of prokaryotes.

One of these facilities is the DOE Joint Genome Institute (JGI) in Walnut Creek, California. The JGI established the program that I lead in Evolutionary Genomics in 2000 to help guide the transition from human genome sequencing to being a comparative genomics institution. Of course, then, our program includes comparative analysis of whole genome sequences (e.g., Dehal et al. 2002; Dehal and Boore 2005, 2006; http://phigs.org/) but, in addition, also sponsors a wide variety of smaller scale projects (see http://evogen.jgi.doe.gov/top_level/projects.html), all funded by external grant support from the U.S. National Science Foundation (MCB-0242131, EAR-0342392, DEB-0120709, DEB-0089624, EF-0228729, EF-0328516, DEB-0416628, DEB-0445047, IOB-0431717, EF-0333173, DBI-0421630, OCE-0313708, DBI-0310028), the U.S. National Institutes of Health (5R01DK066288-02, 1F32GM067463-01), or the U.S. Department of Agriculture (VAR-2002-04334; and Phakopsora genome sequencing direct funding).

Several of these projects focus on sequencing plastid genomes from a wide phylogenetic span of organisms. Plastids are descended from cyanobacteria that took up residence in an early eukaryote (Gray 1988) and gave rise to three lineages: green algae (from which multicellular plants evolved), red algae, and glaucophytes (Moreira, Le Guyader, and Philippe 2000). Subsequently several lineages acquired their plastids secondarily by engulfing either a green or red algae and co-opting the plastid (Delwiche 1999), in some

cases even retaining the engulfed nuclear genome as another organelle termed the nucleomorph (Cavalier-Smith 2002). During this process, this endosymbiotic organism lost many genes, some of which were transferred to form a large component of the nuclear genome (Martin et al. 2002; Raven and Allen 2003; Timmis et al. 2004).

Plastids all retain genomes and are present in plants and many protist groups. Complete sequences are in GenBank for 47 organisms, 35 of which are from Viridiplantae (plants plus related algae), and many more have been sequenced and will soon be available. Sizes for Viridiplantae plastid genomes range from 116,866 bp (*Pinus*) to 203,828 bp (*Chlamydomonas*) and numbers of annotated genes from 128 (*Calycanthus*) to 211 (*Chlorella*), except for the degenerated plastids of non-photosynthetic plants that can be much smaller, with many genes lost (dePamphilis and Palmer 1990; Wolfe, Morden, and Palmer 1992). For the 12 protists with plastid genome sequences available, the size range is from 34,750 bp (the alveolate *Eimeria*) to 183,883 bp (the rhodophyte *Gracilaria*) and the gene count from 63 (the alveolate *Toxoplasma*) to 252 (the rhodophyte *Porphyra*); however, there are a few isolated exceptions known where plastid genomes can be very different, as in dinoflagellates (Zhang, Green, and Cavalier-Smith 1999). Rates of sequence change are generally slow, although genes rearrange commonly in some lineages, mostly by inversions (Cosner et al. 1997; Kim and Lee 2005). For many, there is a large inverted repeat that separates what are called the "large single-copy region" from the "small single-copy region". Introns are occasionally present and post-transcriptional RNA editing can be extensive (Kugita et al. 2003).

Some of our other projects focus on sequencing mitochondrial genomes, also descended from an endosymbiotic prokaryote, in this case an alpha-proteobacteria (Gray 1988; Lang et al. 1997). These genomes are similarly much diminished (Boore 1999; Gray et al. 1998), and also have contributed a great many genes to the eukaryotic nucleus (Lang, Lavrov, and

Burger 2004). All but a few groups of eukaryotes – diplomonads, parabasalads, entamoebae, and microsporidia – contain mitochondria, and a long-standing hypothesis has been that these groups diverged before the endosymbiosis. However, the presence of nuclear genes that appear to have been transferred from a bacterium now supports the view that mitochondria were secondarily lost in these groups (Roger 1999), or perhaps that the hydrogenosomes present are actually homologous to mitochondria (Boore and Fuerstenberg 1999). Why all groups retain at least a vestigial genome in their mitochondria (other than in hydrogenosomes) is still debated.

GenBank contains complete sequences for 764 mtDNAs, 679 of which are from animals. Even within this latter group the taxonomic sampling is highly biased, with 625 being either chordates (504), arthropods (101), or mollusks (20), leaving only 54 for all of the remaining animal phyla. Based on public statements by various investigators, it appears that nearly 1,000 additional complete mtDNA sequences have been determined and are not yet published. Further, there are more than 5,000 entries in GenBank of partial mtDNA sequences that contain at least three contiguous genes. Sizes of those completely sequenced range from 5,957 bp for *Plasmodium* to 430,597 bp for *Nicotiana*, although some not yet sequenced from plants are known to be several megabases in size, similar to bacterial genomes. The number of annotated genes ranges from three for *Plasmodium* to 183 for *Nicotiana*, although many in the latter case are annotated as "hypothetical". Animals, in particular, seem to have conserved an oddly narrow range of variation in both genome size and gene content and a generally slow rate of gene rearrangement (Boore 1999). Rate of sequence change is generally higher in mtDNAs than in nuclear DNA; the most notable exception is plant mtDNAs, which have the lowest rate of evolutionary sequence change of any genome studied (Knoop 2004, but see an exception in Cho et al. 2004).

Much study has been devoted to the molecular biology of mitochondrial systems (Shadel and Clayton 1997; Clayton 2003). Mitochondria import many proteins from the cytoplasm and maintain their own systems for transcription and translation of mitochondrially-encoded genes on mito-ribosomes. They initiate protein synthesis with formyl-methionine as do their prokaryotic progenitors (Smith and Marker 1968). In some cases it is known that genes are transcribed as a polycistron with later enzymatic cleavage of the transcript generating gene specific (or in some cases bicistronic) messages (Battey and Clayton 1980; Ojala et al. 1980). Several modifications of the genetic code are common, including the use of an expanded set of initiation codons (Wolstenholme 1992). In some cases in animal mtDNAs, the cleavage of the polycistron generates an mRNA that ends on a T or TA such that it depends on post-transcriptional polyadenylation to create a TAA stop codon (Ojala et al. 1980); these are called "abbreviated stop codons". Several of these features complicate accurate gene annotation.

Organelle genomes are of interest for a variety of reasons: (1) They form the basis for understanding the evolutionary movement of genes among intracellular compartments (i.e., mitochondrion, plastid, nucleus) (Nugent and Palmer 1991; Daley et al. 2002; Adams and Palmer 2003); (2) Their biochemistry is relatively well understood and includes some of life's most important processes such as ATP production and photosynthesis; (3) Their sequences are commonly used for phylogenetics (e.g., Leebens-Mack et al. 2005; Parham et al. 2006), forensics (e.g., Budowle et al. 2003), population genetics (e.g., Pakendorf and Stoneking 2005), and biogeography (e.g., Macey et al. 2005); (4) Their sequences must co-evolve (e.g., Dey, Barrientos, and Moraes 2000) with the hundreds of nuclear-encoded proteins that are imported and interact, often so intimately as to form multi-subunit enzyme complexes; (5) They provide a suite of genome-level characters such as the relative arrangement of genes for reconstructing ancient evolutionary relationships (Boore and Brown

1998); (6) More generally, they are a model system of genome evolution, where one can investigate changes in protein and RNA secondary structures, in transcription, translation, and replication, in the effect of mutational bias or error correction mechanisms on molecular sequence change, in the patterns of gene rearrangement or stability, and many other genomic features in a relatively simple system.

The pace of sequence production has now outstripped the ability of investigators to make broad comparisons across all of this dataset of complete organellar genomes. There is an urgent need to build a system of databases and query tools that enables comparisons of many features for many organisms. Some efforts have been made (Table 1), but in general they are incomplete and contain many errors.

NCBI (i.e., GenBank) contains nearly all of the sequenced organellar genomes and the corresponding gene annotations, and all other databases are to some extent derivative of these records. There are several very useful features available from the organelle genomes page (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html), including the ability to visualize gene arrangements simultaneously for many taxa and to sort entries taxonomically. There are links to the original files and to the NCBI taxonomy database. The presentation is based on their program called "Refseq" in which they purport to correct and standardize these entries and then assign a new accession number to each curated file. However, in practice, even these entries are replete with error. Even a casual inspection easily identifies many obvious errors, as in these examples: (1) Many files have all of the genes designated on one strand even when some subset of genes should be marked as reverse-complement, as for nine genes in Refseq NC_006295. Even though gene order has almost no variation among vertebrates and these nine genes are on the opposite strand in all other sequenced vertebrate mtDNAs, no attention was given to this in the curation for the Refseq program. (2) Some files contain obviously extraneous gene designations, such as a second gene for tRNA-pro that is only

three nucleotides long in Refseq NC_006899 or a tRNA-glu that is 12,712 nucleotides long in Refseq NC_005280. (3) Some gene designations are missing even when they can be easily found. For example, there is no gene annotation for the small subunit ribosomal RNA (*rrnS*) in Refseq NC_002354, even though it is easily found by a similarity search to be in the unannotated region between *trnC* and *cox2* and even though suspicion might have been aroused when seeing it missing, since it is universally present in all other animal mtDNAs. (4) There are two codon families for each of the amino acids leucine and serine, and so in each case there are two different tRNAs. These identities are erroneously switched in some records, such as in Refseq NC_002544, where the leucine tRNAs say "codons recognized: UUR" and "codons recognized: CUN" in opposite to what is correct, even though a screen of the tRNA anticodons makes this error obvious. (5) There is no consistency in gene names, for example, as when the Refseq just mentioned, NC_002354, uses gene names *ND6* and *CYTB*, whereas others, such as Refseq NC_001637, use *nad6* and *cob* for the same two genes.

As bad as these obvious errors are, the set of non-curated, partially determined sequences is far worse. For example, a large number have the nonsensical gene annotation "tRNA-Asx", whereas there would be no such ambiguity between tRNA-Asn and tRNA-Asp with the genome sequence in hand. In no case is any attention paid to conventions regarding upper vs. lower case lettering for gene names. Oddly, tRNA-encoding genes do not have gene designations and are named for their products (e.g., tRNA-cys) instead of using actual gene names (e.g., *trnC*) (except for occasional inconsistencies, such as in Refseq NC_001807, wherein some tRNA-encoding genes are properly named whereas others are not, and some have gene designations whereas others do not).

As the literature on organelle genomes has become ever larger, and especially as it becomes less common for genomic features to be described in detail in publications, researchers must depend on databases for information. Yet if an analysis depended on the

current set of organelle genome sequences, the results would be grossly in error, with conclusions to be drawn for scores or hundreds of gene gains and losses and transpositions that have not occurred. This current database is at best of little use, and more troubling, it is setting a trap for drawing grossly inaccurate conclusions. This is somewhat ameliorated by the lack of standardization of gene designation and naming conventions, because actually retrieving the erroneous information for such comparisons using scripted search tools is so difficult.

As can be seen in Table 1, there are a few other efforts. The Organellar Genome Retrieval System at McMaster University includes only animal mtDNAs, but allows retrieval of several useful types of information. It does not significantly curate the annotations of GenBank and its gene order comparison tools identify only those that match exactly to a query rather than to a subset of the query gene arrangement. GoBase, the Organellar Genome Database in Canada is a sophisticated attempt at curating the errors in other records and presenting information to the community and it contains information on RNA structures and on RNA editing. A survey of the GenBank errors used as examples above finds that some are corrected (as for the *rrnS* designation in NC_002354 and eight of the nine [all but for the tRNA-glu gene] of the reverse complement designations for NC_006295) but others are not (as for the three nucleotide tRNA gene in NC_006899 and the misassignment of the two leucine tRNAs in NC_002544). Databases at Penn State are offering standardization and curation of completely sequenced plastid genomes. Lastly, the Organelle Genomics website that we are building at the JGI includes a curated set of all animal mtDNAs that are completely sequenced or have at least three contiguous gene sequences and all completely sequenced plastid genome sequences. Automated scripts gather information from GenBank files that are subject to expert curation and standardization of naming conventions. A tool (DOGMA) is provided for semi-automated gene annotation of either organelle (Wyman,

Jansen, and Boore 2004), although it is not yet useful for non-animal mtDNAs. One can browse and search for gene arrangement identities and similarities along with associated information. Although scores of errors have been found and corrected in our databases, others certainly remain, and the pace of new sequencing is overwhelming our ability to continue doing this as a sideline to our main research programs.

So what should be done?

1. *Standardize gene names*. Although experienced researchers may be able to recognize synonymous gene names, newcomers have difficulty, and scripts used to query databases completely fail unless the author properly anticipates and encodes all variations. Any standard would be better than the current system, but the best is to use the names of bacterial homologs in acknowledgement of the prokaryotic past of organelles, as is recommended in Lang et al. (1997), Martin et al. (2002) and Boore, Medina, and Rosenberg (2004) and found at GoBase, DOGMA, the Chloroplast Genome Database, and the Organelle Genomics site at the JGI (see Table 1 for URLs) (but not used consistently in the literature or in other databases). Also, as is conventional, gene names should be in lower case (reserving upper case for their products) and tRNA-encoding genes should have the form *trnX*, where X is the one-letter code for the corresponding amino acid.

2. *Label tRNAs with anticodon*. Homologies are best indicated by appending the anticodon in parentheses to the name of tRNA-encoding genes, for example *trnL(taa)* or *trnS(tct)*. Although it is common in publications to use codon recognized, this is one step more inferential and could potentially be later shown by experiment to be inaccurate, whereas the anticodon is apparent from the genome sequence. (The convention is to use upper case for codon, lower case for anticodon.) One must remember, of course, that codons of

even homologous tRNA genes may vary, as in *trnK(ctt)* and *trnK(ttt)*, only one of which is found for each animal mtDNA, so another possibility is to designate homology with the maximally ambiguous anticodon, in this case *trnK(ytt)*.

3. *Standardize the format for designating genes*. Database queries are frustrated by having inconsistencies such as some files where tRNA-encoding genes are labelled "tRNA" and others are labelled "gene".

4. *Establish standards for designating gene boundaries*. Gene annotation is complicated by the variety of alternative start codons (Wolstenholme 1992), the use of incomplete stop codons completed by polyadenylation (Ojala et al. 1980), and post-transcriptional RNA editing (Lavrov, Brown, and Boore 2000; Gray 2003; Kugita et al. 2003). Consequently, inferring the exact beginning and end of genes from genome sequence alone is sometimes ambiguous. Several factors must be considered, including the commonality of these variations in related organisms, the degree of similarity of gene predictions with homologs, and the possibility of gene overlap. Correction of submitted annotations should enforce a standard of inference so that comparisons among these genomes are valid.

5. *Establish standards for accepting the reality of a gene assignment.* Open reading frames (ORFs) are found by chance in DNA sequence with a calculable probability. Some represent actual genes and others do not. The best evidence from sequence alone is to be found in the conservation of an ORF across a phylogenetic distance where sufficient sequence change has occurred to expect ORF disruption, indicating that purifying selection has been operating. We should establish standards for designating ORFs as genes that clearly differentiate the strength of evidence. Names should indicate identified homologies.

6. *Perform systematic error screening*. Automated scripts should be built that screen for indicators of likely errors of annotation so that these files can be passed to an expert curator. For example, one might find organelle genomes that are most similar in sequence and identify any gene losses or rearrangements that might indicate a misannotation.

7. *Include information on RNA editing*. Codified in the sequence files should be information on known or inferred RNA edit site such that these can be visualized and extractions can be made with either edited or unedited sequences.

8. *Automate sequence alignments and phylogenetic analyses*. Automated scripts should be used to extract sequences of individual genes, align them, and construct phylogenetic analyses, as is being done for limited subsets of taxa at the Chloroplast Genome Database and at the Mammalian Mitochondrial Genomics Database (http://www.mammibase.lncc.br/).

9. *Add more descriptors*. Sequence files should commonly include information such as collection locality, location of a voucher specimen (ideally with a museum accession number), and much more information on the procedures used to determine the sequence, since much of this will never be included in any publication.

10. *Fully integrate this information with data from prokaryotic genomes*. Plastids and mitochondria descended from cyanobacteria and alpha-proteobacteria, respectively, and in so doing, contributed a very large number of genes to the eukaryotic nucleus. A comprehensive database should be built that reconstructs the grand sweep of this evolution, starting with the endosymbioses and leading us to the distribution of genes in modern organisms, whether currently resident in the nucleus or in organelles, by including the homologous genes in prokaryotes. An aid to this may be in the scripts we've written for "PhIGs", or "Phylogenetically Inferred Groups", a system for

automating accurate gene family construction, phylogenetic analyses, and interpretation and presentation of results.

Little funding has been made available for directly addressing these needs and the tasks are now overwhelming the ability of these few investigators to do this as a sideline. With focused effort, this could well develop into a set of databases and query tools that not only addresses a series of interesting biological questions, but also forms a model for database development for larger genomes.

ACKNOWLEDGEMENTS

Thanks to Dawn Field, Peter Sterk, and others at EBI for their inspiration and energy in organizing these efforts. This work was supported by NSF grants EAR-0342392, DEB-0120709, DEB-0089624, EF-0228729, EF-0328516, and DEB-0416628, and was performed partly under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231.

REFERENCES

ADAMS, K.L., and PALMER, J.D. (2003). Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. Mol. Phylogenet. Evol. 29, 380-395.

BATTEY, J., and CLAYTON, D.A. (1980). The transcription map of human mitochondrial DNA implicates transfer RNA excision as a major processing event. J. Biol. Chem. 255, 11599-11606.

BOORE, J.L. (1999). Animal mitochondrial genomes. Nucleic Acids Res. 27, 1767-1780.

BOORE, J.L., and BROWN, W.M. (1998). Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. Curr. Opin. Gen. Dev. 8, 668-674.

BOORE, J.L., and FUERSTENBERG, S.I. (1999). *Entomoeba histolytica*—A derived mitochondriate eukaryote? Trends Microbiol. 7, 426-428.

BOORE, J.L., MEDINA, M., and ROSENBERG, L.A. (2004). Complete sequences of two highly rearranged molluscan mitochondrial genomes, those of the scaphopod *Graptacme eborea* and of the bivalve *Mytilus edulis*. Mol. Biol. Evol. 21, 1492–1503.

BUDOWLE, B., ALLARD, M.W., WILSON, M.R., and CHAKRABORTY, R. (2003). Forensics and mitochondrial DNA: applications, debates, and foundations. Annu. Rev. Genomics Hum. Genet. 4, 119-141.

CAVALIER-SMITH, T. (2002). Nucleomorphs: Enslaved algal nuclei. Curr. Opin. Microbiol. 5, 612-619.

CHO, Y., MOWER, J.P., QIU, Y.L., and PALMER, J.D. (2004). Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc. Natl. Acad. Sci. USA 101, 17741-17746.

CLAYTON, D.A. (2000). Transcription and replication of mitochondrial DNA. Hum. Reprod. 15 Suppl 2, 11-17.

CLAYTON, D.A. (2003). Mitochondrial DNA replication: What we know. IUBMB Life 55, 213-217.

COSNER, M.E., JANSEN, R.K., PALMER, J.D., and DOWNIE, S.R. (1997). The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/ deletions, and several repeat families. Curr. Genet. 31, 419-429.

DALEY, D.O., ADAMS, K.L., CLIFTON, R., QUALMANN, S., MILLAR, A.H., PALMER, J.D., PRATJE, E., and WHELAN, J. (2002). Gene transfer from mitochondrion to nucleus: novel mechanisms for gene activation from Cox2. Plant J. 30, 11-21.

DEHAL, P., and BOORE, J.L. (2005). Two rounds of genome duplication in the ancestral vertebrate genome. PLoS Biology 3(10), e314.

DEHAL, P., and BOORE, J.L. (2006). The PhIGs (phylogenetically inferred groups) database: A resource for phylogenomics. BMC Bioinformatics, in press.

DEHAL, P., SATOU, Y., CAMPBELL, R., CHAPMAN, J., DEGNAN, B., and DETOMASO, A. et al. (2002). The draft Genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. Science 298, 2157-2167.

DELWICHE, C.F. (1999). Tracing the tread of plastid diversity through the tapestry of life. Am. Nat. 154(S4), S164-177.

DEPAMPHILIS, C.W., and PALMER, J.D. (1990). Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature 348, 337-339.

DEY, R., BARRIENTOS, A., and MORAES, C.T. (2000). Functional constraints of nuclear-mitochondrial DNA interactions in xenomitochondrial rodent cell lines. J. Biol. Chem. 275, 31520-31527.

GRAY, M.W. (1988). Organelle origins and ribosomal RNA. Biochem. Cell Biol. 66, 325-348.

GRAY, M.W. (2003). Diversity and evolution of mitochondrial RNA editing systems. IUBMB Life 55, 227-233.

GRAY, M.W., LANG, B.F., CEDERGREN, R., GOLDING, G.B., LEMIEUS, C., SANKOFF, D. TURMEL, M., BROSSARD, N. DELAGE, E., LITTLEJOHN, T.G., PLANTE, I. RIOUS, P. SAINT-LOUIS, D., ZHU, Y., and BURGER, G. (1998). Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res. 26, 865-878.

KIM, K.J., and LEE, H.L. (2005). Widespread occurrence of small inversions in the chloroplast genomes of land plants. Mol. Cells 19, 104-113.

KNOOP, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. 46, 123-139.

KUGITA, M., YAMAMOTO, Y. FUJIKAWA, T., MATSUMOTO, T., and YOSHINAGA, K. (2003). RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res. 31, 2417-2423.

LANG, B.F., BURGER, G., O'KELLY, C.J., CEDERGREN, R., GOLDING, G.B., LEMIEUX, C., SANKOFF, D., TURMEL, M., and GRAY, M.W. (1997). An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387, 493-497.

LANG, B.F., LAVROV, D.V., and BURGER, G. (2004). Mitochondrial Genome, Evolution. Encyclopedia of Biological Chemistry, Volume 2. (Elsevier, Inc., London).

LAVROV, D., BROWN, W.M., and BOORE, J.L. (2000). A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede *Lithobius forficatus*. Proc. Natl. Acad. Sci. 97, 13738-13742.

LEEBENS-MACK, J., RAUBESON, L.A., CUI, L., KUEHL, J.V., FOURCADE, H.M, CHUMLEY, T.W., BOORE, J.L., JANSEN, R.K., and DEPAMPHILIS, C.W. (2005). Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. Mol. Biol. Evol. 22, 1948-1963.

MACEY, J.R., FONG, J.J., KUEHL, J.V., SHAFIEI, S., ANANJEVA, N.B., PAPENFUSS, T.J., and BOORE, J.L. (2005). The complete mitochondrial genome of a gecko and the phylogenetic position of the Middle Eastern *Teratoscincus keyserlingii*, Mol. Phylogenet. Evol. 36, 188-193.

MARTIN, W., RUJAN, T., RICHLY, E., HANSEN, A., CORNELSEN, S., LINS, T., LEISTER, D., STOEBE, B., HASEGAWA, M., and PENNY, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. USA 99, 12246-12251.

MOREIRA, D., LE GUYADER, H., and PHILIPPE, H. (2000). The origin of red algae and the evolution of chloroplasts. Nature 405, 69-72.

NUGENT, J.M., and PALMER, J.D. (1991). RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. Cell 66, 473-481.

OJALA, D., MERKEL, C., GELFAND, R., and ATTARDI, G. (1980). The tRNA genes punctuate the reading of genetic information in human mitochondrial DNA. Cell 22, 393-403.

PAKENDORF, B., and STONEKING, M. (2005). Mitochondrial DNA and human evolution. Annu. Rev. Genomics Hum. Genet. 6, 165-183.

PARHAM, J.F., MACEY, J.R., PAPENFUSS, T.J., FELDMAN, C.R., TURKOZAN, O., POLYMENI, R., and BOORE, J.L. (2006). The phylogeny of Mediterranean tortoises

and their close relatives based on complete mitochondrial genome sequences from museum specimens. Mol. Phylogenet. Evol. 38, 50-64.

RAVEN, J.A., and ALLEN, J.F. (2003). Genomics and chloroplast evolution: What did cyanobacteria do for plants? Genome Biol. 4, 209.

ROGER, A.J. (1999). Reconstructing early events in eukaryotic evolution. Am. Nat. 154(S4), S146-163.

SHADEL, G.S., and CLAYTON, D.A. (1997). Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. 66, 409-435

SMITH, A.E., and MARKER, K.A. (1968). *N*-Formylmethionyl transfer RNA in mitochondria from yeast and rat liver. J. Mol. Biol. 38, 241-243.

TIMMIS, J.N., AYLIFFE, M.A., HUANG, C.Y., and MARTIN, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5, 123-135.

WOLFE, K.H., MORDEN, C.W., and PALMER, J.D. (1992). Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA 89, 10648-10652.

WOLSTENHOLME, D.R. (1992). Animal mitochondrial DNA: Structure and evolution. Intl. Rev. Cytology 141, 173-216.

WYMAN, S., JANSEN, R.K., and BOORE, J.L. (2004). Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252-3255.

ZHANG, Z., GREEN, B.R., and CAVALIER-SMITH, T. (1999). Single gene circles in dinoflagellate chloroplast genomes. Nature 400, 155-159.

Table 1. Databases for organellar genomics

Name	Main features	URL
Organelle Genome Resources at GenBank	Whole genome comparisons of both organelles	http://www.ncbi.nlm.nih.gov/genomes/organelles/organelles.html
Organellar Genome Retrieval System	Information on animal mt gene order and codon usage. Database of sequences for retrieval. Individual gene alignments.	http://drake.physics.mcmaster.ca/ogre/
GoBase, Organelle Genome Database	Curated information on both organellar genomes	http://megasun.bch.umontreal.ca/gobase/gobase.html
MitoDat, Mendelian Inheritance and the Mitochondrion	Nuclear encoded genes producing products that function in mitochondria	http://www-lecb.ncifcrf.gov/mitoDat/
DOGMA, Dual Organellar GenoMe Annotator	Tools for gene annotation of mtDNAs and cpDNAs	http://evogen.jgi.doe.gov/dogma
Chloroplast Genome Database	Curated annotations for whole plastid genome sequences	http://chloroplast.cbio.psu.edu/
Organelle Genomics at DOE Joint Genome Institute	Comparisons of a curated set of all organelle genome sequences both complete and partial. Gene order comparisons.	http://www.jgi.doe.gov/programs/comparative/top_level/organelles.html