D692–D696 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj055

ChloroplastDB: the Chloroplast Genome Database

Liying Cui, Narayanan Veeraraghavan[1], Alexander Richter[1], Kerr Wall, Robert K. Jansen[2], Jim Leebens-Mack, Izabela Makalowska[1] and Claude W. dePamphilis*

Department of Biology and Institute of Molecular Evolutionary Genetics, [1]Center for Computational Genomics, Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA and [2]Section of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA

Received August 19, 2005; Revised and Accepted October 4, 2005

ABSTRACT

The Chloroplast Genome Database (ChloroplastDB) is an interactive, web-based database for fully sequenced plastid genomes, containing genomic, protein, DNA and RNA sequences, gene locations, RNA-editing sites, putative protein families and alignments (http://chloroplast.cbio.psu.edu/). With recent technical advances, the rate of generating new organelle genomes has increased dramatically. However, the established ontology for chloroplast genes and gene features has not been uniformly applied to all chloroplast genomes available in the sequence databases. For example, annotations for some published genome sequences have not evolved with gene naming conventions. ChloroplastDB provides unified annotations, gene name search, BLAST and download functions for chloroplast encoded genes and genomic sequences. A user can retrieve all orthologous sequences with one search regardless of gene names in GenBank. This feature alone greatly facilitates comparative research on sequence evolution including changes in gene content, codon usage, gene structure and post-transcriptional modifications such as RNA editing. Orthologous protein sets are classified by TribeMCL and each set is assigned a standard gene name. Over the next few years, as the number of sequenced chloroplast genomes increases rapidly, the tools available in ChloroplastDB will allow researchers to easily identify and compile target data for comparative analysis of chloroplast genes and genomes.

INTRODUCTION

As the site in the eukaryotic cell where photosynthesis takes place, chloroplasts are responsible for much of the world's primary productivity, making chloroplasts essential to the lives of plants and animals alike. The oxygen in our atmosphere, all agricultural commodities and fossil fuels such as coal and oil are 'products' of photosynthesis (1). Other important activities that occur in chloroplasts (and several types of non-photosynthetic plastids) include the production of starch (2), certain amino acids and lipids (3,4), some of the colourful pigments in flowers (5), and key aspects of sulfur and nitrogen metabolism (6,7).

All plastids studied to date contain their own distinct genomes derived from a cyanobacterial ancestor that was captured early in the evolution of the eukaryotic cell (8). Although much smaller than the nuclear genome, chloroplast genomes typically contain ~110–120 unique genes including conserved open reading frames (ORFs) annotated as *ycf* genes (hypothetical chloroplast ORF) (9). Additional possible coding regions are designated as ORFs. These are typically annotated with the number of amino acids encoded (e.g., ORF1995) (10). Some algae have retained a large chloroplast genome with >200 genes, whereas the plastid genomes from non-photosynthetic organisms may retain only a few dozen genes.

Chloroplast gene sequences have been widely used as genetic markers for plant and algal phylogenetic studies for nearly two decades (11,12). Whereas one or a few genes have been the focus of study most of this time [*rbc*L, *atp*B, *mat*K; but see studies by Graham and Olmstead (13,14)], rapid growth in the number of chloroplast genome sequences is now making it possible for a wide range of phylogenetic issues to be addressed with genome scale datasets (15,16). For population-level studies, polymorphic regions for targeted sequencing can be identified through comparison of complete genome sequences for exemplar taxa (17). Chloroplast genome sequences are also being used to address a wide range of evolutionary questions about changes in gene content and gene order (18), the dynamics of insertion and deletion events (19), intergenomic gene transfer (20) and photosynthetic evolution (21). The development of genetic transformation of chloroplasts has been very exciting (22) and the list of target species will increase as the locations and flanking sequences for intergenic spacer regions are identified from an expanding number of chloroplast genome sequences (23). Genome-scale

*To whom correspondence should be addressed. Tel: +1 814 863 6412; Fax: +1 814 865 9131; Email: cwd3@psu.edu

functional analyses, including investigations of plastid transcriptomes and proteomes are also progressing rapidly (24).

Several bioinformatic resources provide information on organelle genomes, and tools specific for these genomes have been developed (25). The standard repository for full genome sequences, the GenBank, EMBL and DDBJ nucleotide sequence databases, currently includes 44 complete plastid genomes sequenced since 1986. The NCBI GenBank genome section lists entire organelle genome sequences submitted to the database and reviewed by NCBI staff (26). GOBASE (27) also maintains a list of sequenced organelle genomes. A standardized nomenclature for plastid-encoded protein genes is available through the UniProt database (http://www.expasy.org/txt/plastid.txt). A web-based annotation tool, DOGMA, provides a graphical user interface to annotate draft and finished organelle genomes based on sequence similarity searches and RNA secondary structure prediction (28). The program GRAPPA has been used for phylogenetic analysis of chloroplast gene order changes (29). A plastid gene order database was developed with uniform gene names for 32 plastid genomes (30). In addition, 500 primers are now available for targeted PCR amplification of sequences from chloroplast genomes (http://bfw.ac.at/200/1859.html).

A prerequisite of future research is the accessibility of well-annotated, easy-to-use sequence data. However, several major limiting factors exist including flat file presentation of annotated organelle genomes, lack of standard data structure for relational databases, and non-uniform annotation quality. Errors in the annotation typically persist in the standard databases (e.g. the gene *rpl2* is annotated as *rp12* in the *Oryza sativa* chloroplast DNA, a human error). As a heritage of early annotations, gene name variants, unidentified *ycf*s and ORFs, and unannotated genes are present in some genomes. Given the ubiquity of phylogenetic studies based on plastid gene sequences, the flat file format makes search and data retrieval cumbersome.

RNA editing, a post-transcriptional process that alters specific RNA bases prior to translation, is common in the chloroplast genomes of some land plants (31). RNA editing can result in the creation of start codons and removal of stop codons, as well as making radical amino acid substitutions that would not be predicted based on the DNA sequence alone. Accurate genome annotation and inference of protein sequences often cannot be accomplished without knowledge of RNA editing sites (e.g. in the chloroplast DNA of *Anthoceros formosae*, *Adiantum capillus-veneris* and *Zea mays*). The pace of new data generation and large-scale analyses demand a better integration of resources for chloroplast genome research. ChloroplastDB is a relational database with a user-friendly interface and tools to aid the analysis of chloroplast genome sequences.

DATA MANAGEMENT AND ORGANIZATION

ChloroplastDB was designed using a MySQL database structure (Figure 1). The tables in the relational database store data related to the genes, nucleotide sequences and annotated protein sequences for coding sequences (CDS). The databases contain fully sequenced plastid genomes obtained from the NCBI RefSeq section (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html). All genes, including protein-coding genes, tRNA, rRNA, hypothetical ORFs (*ycf*, ORF) were parsed and incorporated into the database.

The standard process for extracting and storing data was carried out as follows:

(i) A GenBank XML file containing a plastid genome sequence is downloaded. The XML format ensures better integrity of parsed data than GenBank flat files.
(ii) Using in-house XML parsers written in Perl, the XML data is extracted, filtered through quality control steps and formatted properly. The cleaned data are stored in the database in a form conducive to efficient data transactions.
(iii) Using the coordinates from the features (CDS, tRNA, rRNA, intron), the corresponding nucleotide sequence is extracted from the genome and stored in the database.
(iv) In a few instances when a parsed sequence lacks appropriate annotation, the GenBank records are updated with expert annotations after automatic processing of the XML file.
(v) Three BLAST databases are created: one for the whole genome sequence from all organisms, a second for the annotated protein sequences of all organisms in the database, and a third for the generated nucleotide CDS from each organism.

When new sequences are added to the database, the proteins are sorted into potentially orthologous sets or 'tribes' using tribeMCL (32). First, a sequence similarity profile is obtained by all-against-all BLAST on the protein sequences at a threshold of 1E-3. The BLAST output is fed to tribeMCL, which then generates a list of tribes representing protein families. This output is parsed and the tribes are updated in the database.

The quality control procedure is crucial in maintaining the integrity and accuracy of the extracted data. There appear to be some irregularities with the GenBank annotations. The genomic region spanning one gene and the gene features (CDS, tRNA, rRNA) share the same gene name. In case of overlapping and nested genes, the annotation for the second (or nested) gene could be attributed to the first gene, resulting in confusions. Also, there are instances where the gene names are not included in the feature description. We have avoided the problem by using the coordinates for each 'gene' feature as the primary reference, and after that, the coordinates of other features are checked and assigned new gene names. For example, *rps12* is a *trans*-spliced gene containing three exons in the angiosperm chloroplast genome. The first exon is located ~30 kb upstream of the second exon, and on the opposite coding strand. The initial parsed record for the gene contained many other genes nested in the intron region. After the filtering step, those nested genes were dissociated from the name '*rps12*' and assigned to their appropriate names. If no gene name was found, the feature was deemed to be an 'ophan' and assigned a local name (starting with 'lcl_anno').

When the GenBank annotation included RNA edited sites, both the location and type of edits were extracted from the record. The information was used to generate an edited pseudosequence that was stored with a list of edited sites. Just 38 genes with 541 annotated, experimental verified sites from *A.formosae* and *Physcomitrella patens* are included in the current GenBank annotations. RNA editing has been reported

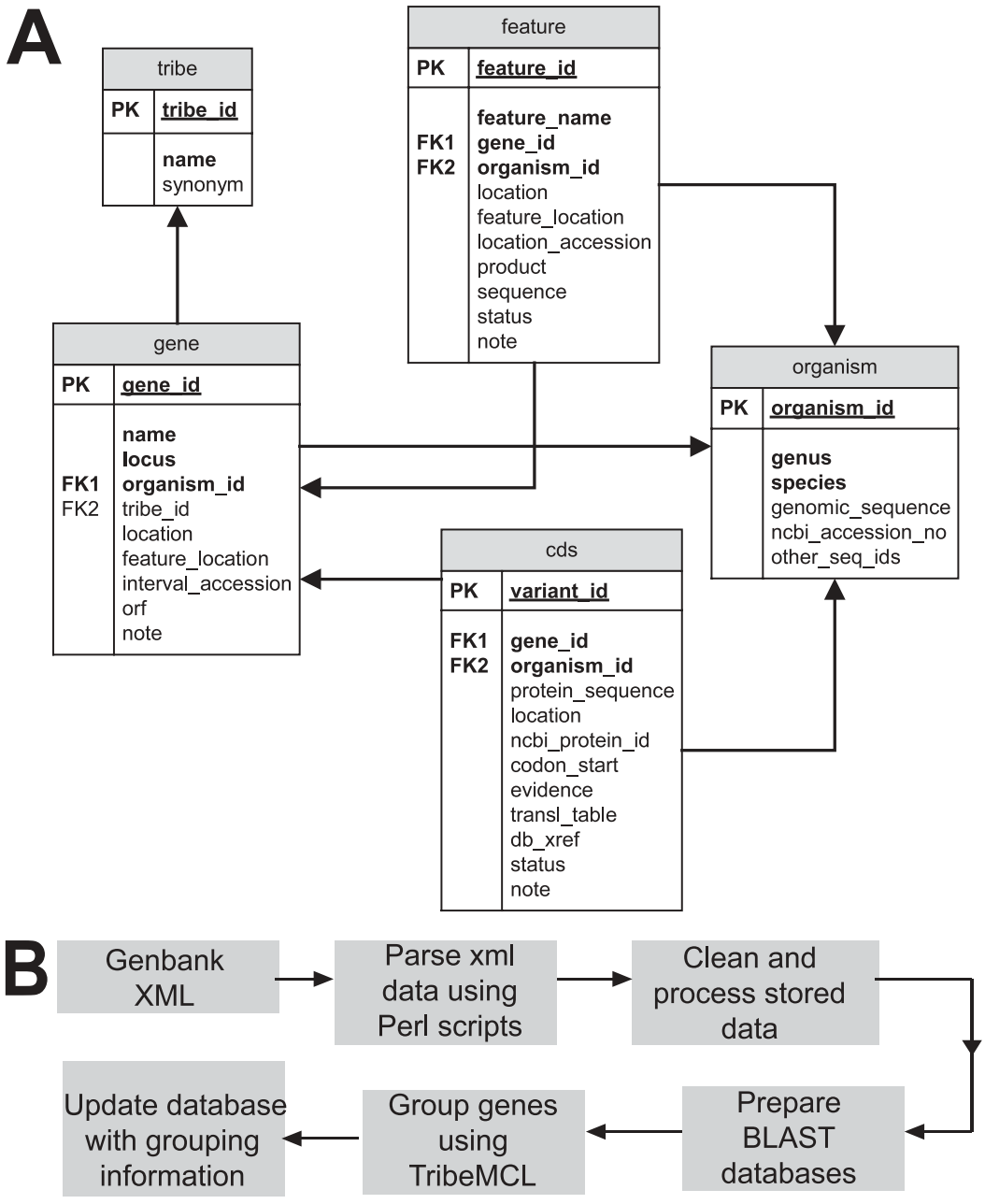

Figure 1. ChloroplastDB overview. (**A**) Database structure and relationship of data tables. PK: primary key. FK: foreign key. (**B**) Data flow and filtering steps to ensure the high quality of data stored in the database.

in other plants including tobacco, maize and *Adiantum* chloroplast DNA. Because the GenBank record does not contain a standard feature to store the RNA editing information, some edited sites were not reported while others were reported as exceptions since the protein sequence did not match conceptual translation of the protein coding gene. To maintain quality and consistency of the data, we report annotated locations and the edited mRNA sequence.

THE CHLOROPLAST GENOME DATABASE INTERFACE

Web user interfaces, developed using Perl CGI scripts, interact with the above mentioned data repository and provide users with basic sequence analysis tools (Figure 2). ChloroplastDB can be queried by gene name, and query results are returned in a table with links to individual genes. The BLAST similarity search was implemented for search against whole genome, extracted proteins or extracted CDS. Sequences returned in BLAST searches can be exported to a fasta file. A user can also browse the list of organisms and all genes by specified subtypes (tRNA, rRNA and protein-coding) from each organism. The set of extracted genes vary from 56 in a non-photosynthetic parasitic plant, *Epifagus virginiana*, to 254 in the red alga *Porphyra purpurea*, including duplicate genes that are present in the genome. Tribes represent putatively orthologous genes across organisms, which can be downloaded to construct multiple sequence alignments. Together,

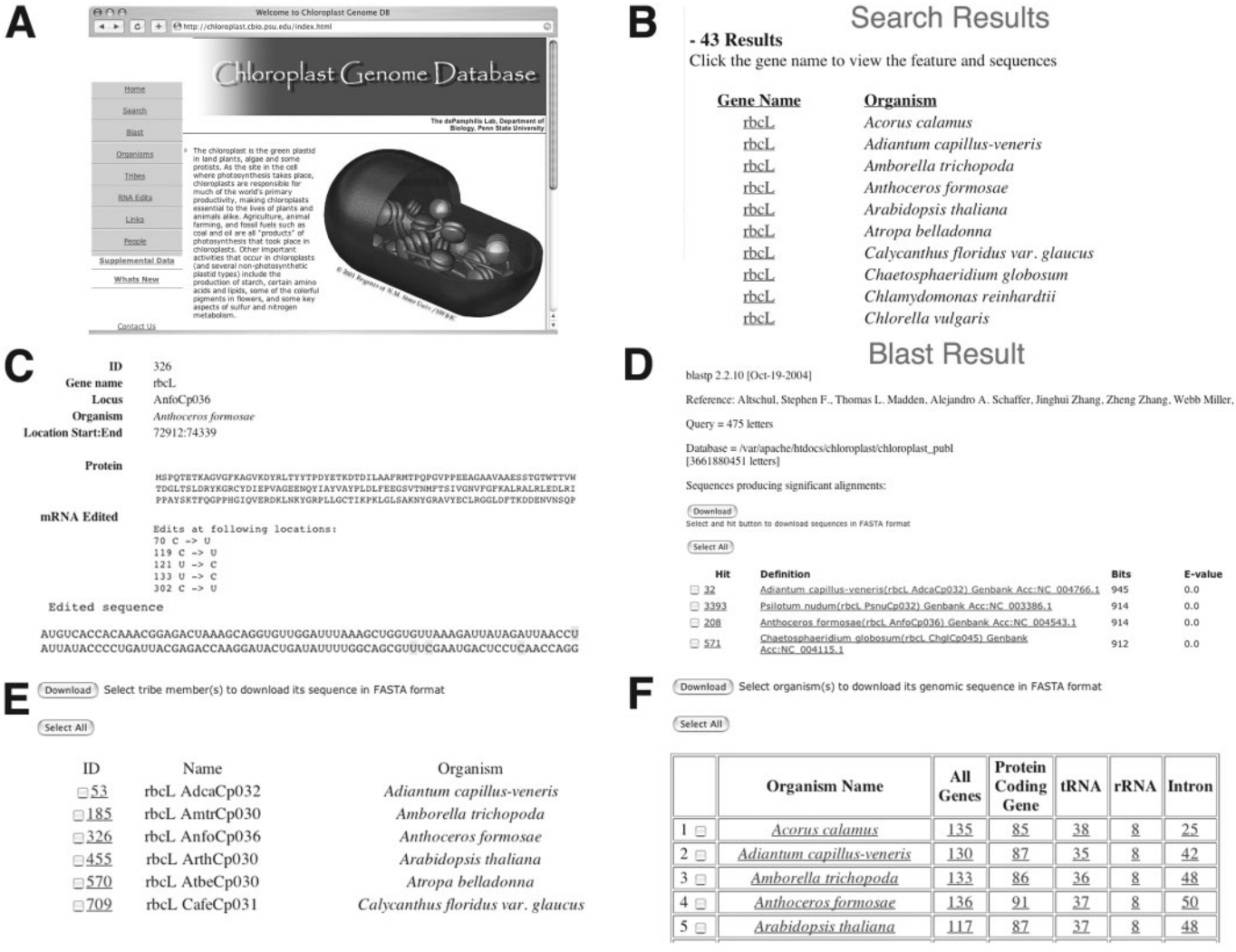

Figure 2. Examples of analysis using the ChloroplastDB web interface. (**A**) Homepage of the database. (**B**) Search results for the gene '*rbc*L'. (**C**) The gene view page linked to search result for each gene, including mRNA editing information. (**D**) BLAST results, with options to download sequences from the BLAST search. (**E**) Putative orthologous gene set listed as 'Tribes'. (**F**) The organism page presents a summary of genomes and extracted features in the database for batch download.

these web interfaces provides a workbench for query, search, and sequence compilation and analysis. The various functions are seamlessly linked for a smooth user experience.

HOW TO USE THE CHLOROPLAST DATABASE

Gene search

The basic query page allows a user to search for individual gene of interest. For example, search of '*rbc*L' returns all *rbc*L gene entries, in which two copies are from *Nephroselmis olivacea* since they are duplicated and located in the inverted repeats. Each gene is then linked to a gene view page. The gene view displays the gene name, organism, coordinates on the genome, exon boundaries, and DNA or protein sequences. Annotated RNA edits are highlighted with colours for easy identification.

BLAST

Customized BLAST searches against nucleotide CDS, proteins or genomic sequences allows a researcher to quickly identify novel sequences, to construct alignments and to annotate chloroplast genes. The returned entries are linked to respective gene view page or the whole genome record in NCBI. Selected list of entries can be exported as fasta sequences. A user can also run BLAST against the *Arabidopsis* or rice proteome to identify nuclear encoded homologs of chloroplast genes.

Tribes

An important feature of this database is pre-computed orthologous protein sets, which could be used for phylogenetic analysis. The tribes present a uniform, automatic classification of chloroplast proteins using MCL clustering on all-by-all BLAST search results. With few exceptions, all other tribes represent orthologous gene sets, and a standard name is displayed for each tribe according to the UniProt list of plastid and cyanelle genes. The paralogous *psa*A and *psa*B are highly similar duplicate genes (BLAST E-value < 1.0E−150) which are grouped together in a single tribe. In contrast, rapidly evolving *ycf1* genes are split into three tribes including seed plant, ferns plus bryophytes and algal orthologues. Tribes

also become a discovery tool for unannotated proteins. For example, ORF288 in hornwort, *A.formosae*, was sorted to the *cys*T tribe, together with an unannotated orthologous sequence from liverwort, *Marchantia polymorpha*. We also provide precomputed protein and DNA alignments for each tribe.

Whole genome comparison and batch sequence retrieval

The plastid genomes from land plants, green algae, red algae and Apicomplexian represents a great range of diversity of the organelle genomes. The organism page presents direct link to the GenBank genome sequences, and ability to download genome sequences and genes by organisms. The user can use the downloaded sequence for organism specific analysis, or comparison for a specific type of sequences across organisms.

FUTURE PROSPECTS

Over the next few years, the growth of full organelle genome sequences will provide new opportunities for whole-genome comparative analyses. Cross-species investigations of genome-wide structural evolution, context-specific substitution processes (33), RNA editing, gene regulation and gene function will be more tractable for organelle genomes than much larger and more complex nuclear genomes. Organelle genomes may be an ideal proving ground for methods of analysis being developed to understand genome and gene order evolution. The mission of ChloroplastDB is to promote comparative analyses of plastid genomes by addressing the community need for better, uniform annotation, quick sequence retrieval and homology search tools. The functionality of ChloroplastDB will grow as new genomes, alignments and other analyses are added, gene clustering techniques are improved, and visualization tools with gene order browsers are developed.

ACKNOWLEDGEMENTS

The authors thank Kevin Beckmann and Stacia Wyman for programming support, and Paul Wolf for providing RNA editing information. This work was supported through NSF grants DBI 01-15684 to C.W.D. and DEB 01-20709 to R.K.J. and C.W.D., and Eberly College of Science, Pennsylvania State University. Funding to pay the Open Access publication charges for this article was provided by NSF DEB 01-20709.

Conflict of interest statement. None declared.

REFERENCES

1. Halliwell,B. (1978) The chloroplast at work. A review of modern developments in our understanding of chloroplast metabolism. *Prog. Biophys. Mol. Biol.*, **33**, 1–54.
2. Baroja-Fernandez,E., Munoz,F.J., Akazawa,T. and Pozueta-Romero,J. (2001) Reappraisal of the currently prevailing model of starch biosynthesis in photosynthetic tissues: a proposal involving the cytosolic production of ADP-glucose by sucrose synthase and occurrence of cyclic turnover of starch in the chloroplast. *Plant Cell Physiol.*, **42**, 1311–1320.
3. Kirk,J.T. (1971) Chloroplast structure and biogenesis. *Annu. Rev. Biochem.*, **40**, 161–196.
4. Vothknecht,U.C. and Westhoff,P. (2001) Biogenesis and origin of thylakoid membranes. *Biochim. Biophys. Acta*, **1541**, 91–101.
5. Reinbothe,S. and Reinbothe,C. (1996) The regulation of enzymes involved in chlorophyll biosynthesis. *Eur. J. Biochem.*, **237**, 323–343.
6. Hatzfeld,Y., Lee,S., Lee,M., Leustek,T. and Saito,K. (2000) Functional characterization of a gene encoding a fourth ATP sulfurylase isoform from *Arabidopsis thaliana*. *Gene*, **248**, 51–58.
7. Schiltz,S., Gallardo,K., Huart,M., Negroni,L., Sommerer,N. and Burstin,J. (2004) Proteome reference maps of vegetative tissues in pea. An investigation of nitrogen mobilization from leaves during seed filling. *Plant Physiol.*, **135**, 2241–2260.
8. Margulis,L. (1975) Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symp. Soc. Exp. Biol.*, 21–38.
9. Rochaix,J.D. (1997) Chloroplast reverse genetics: new insights into the function of plastid genes. *Trends Plant Sci.*, **2**, 419–425.
10. Hallick,R.B. and Bairoch,A. (1994) Proposal for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. *Plant Mol. Biol. Rep.*, **12**, S29–S30.
11. Cattolico,R.A. (1985) Chloroplast biosystematics: chloroplast DNA as a molecular probe. *Biosystems*, **18**, 299–306.
12. Clegg,M.T. (1993) Chloroplast gene sequences and the study of plant evolution. *Proc. Natl Acad. Sci. USA*, **90**, 363–367.
13. Graham,S.W. and Olmstead,R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.*, **87**, 1712–1730.
14. Graham,S.W. and Olmstead,R.G. (2000) Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Curr. Genet.*, **37**, 183–188.
15. Leebens-Mack,J., Raubeson,L.A., Cui,L., Kuehl,J.V., Fourcade,M.H., 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.
16. Goremykin,V.V., Holland,B., Hirsch-Ernst,K.I. and Hellwig,F.H. (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Mol. Biol. Evol.*, **22**, 1813–1822.
17. Provan,J., Powell,W. and Hollingsworth,P.M. (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol. Evol.*, **16**, 142–147.
18. Gray,M.W. (1999) Evolution of organellar genomes. *Curr. Opin. Genet. Dev.*, **9**, 678–687.
19. Ingvarsson,P.K., Ribstein,S. and Taylor,D.R. (2003) Molecular evolution of insertions and deletion in the chloroplast genome of silene. *Mol. Biol. Evol.*, **20**, 1737–1740.
20. Martin,W., Stoebe,B., Goremykin,V., Hapsmann,S., Hasegawa,M. and Kowallik,K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162–165.
21. Bungard,R.A. (2004) Photosynthetic evolution in parasitic plants: insight from the chloroplast genome. *Bioessays*, **26**, 235–247.
22. Daniell,H. and Chase,C.D. (eds) (2004) *Molecular Biology and Biotechnology of Plant Organelles*. Kluwer Academic Publishers, Dordrecht.
23. Daniell,H. (1999) New tools for chloroplast genetic engineering. *Nat. Biotechnol.*, **17**, 855–856.
24. Rochaix,J.D. (2001) Posttranscriptional control of chloroplast gene expression. From RNA to photosynthetic complex. *Plant Physiol.*, **125**, 142–144.
25. Jansen,R.K., Raubeson,L.A., Boore,J.L., dePamphilis,C.W., Chumley,T.W., Haberle,R.C., Wyman,S.K., Alverson,A.J., Peery,R., Herman,S.J. *et al.* (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.*, **395**, 348–384.
26. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI reference sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
27. O'Brien,E.A., Badidi,E., Barbasiewicz,A., deSousa,C., Lang,B.F. and Burger,G. (2003) GOBASE—a database of mitochondrial and chloroplast information. *Nucleic Acids Res.*, **31**, 176–178.
28. Wyman,S.K., Jansen,R.K. and Boore,J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.
29. Moret,B.M., Wang,L.S., Warnow,T. and Wyman,S.K. (2001) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17** (Suppl. 1), S165–S173.
30. Kurihara,K. and Kunisawa,T. (2004) A gene order database of plastid genomes. *Data Sci. J.*, **3**, 60–79.
31. Sugiura,M. (1995) The chloroplast genome. *Essays Biochem.*, **30**, 49–57.
32. Enright,A.J., Kunin,V. and Ouzounis,C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.*, **31**, 4632–4638.
33. Morton,B.R. and Clegg,M.T. (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J. Mol. Evol.*, **41**, 597–603.