modern *H. sapiens* (Abri Pataud, Obercassel II); male Neanderthals (La Chapelle aux Saints, La Ferrassie I, Monte Circeo, La Quina V); female Neanderthals (Gibraltar I); male *H. heidelbergensis* (Broken Hill, Petralona); female *H. heidelbergensis* (Steinheim); male *H. erectus* (OH 9); female *H. erectus* (KNMR-ER 3733). I radiographed all crania except for Skhul IV (B. Arensburg), Petrolona (C. Stringer), KNM-ER 3733 (A. Walker) and Obercassel I and II, Monte Circeo, La Quina V (T. Molleson). F. Spoor provided CT scans of the OH 9, Broken Hill and Steinheim specimens.

**Measurements.** Linear and angular measurements were taken from traced radiographs using digital calipers accurate to 0.01 mm, and a protractor accurate to 1°. Measurements include: ASL (anterior sphenoid body length), the minimum distance from the sella to the posterior maxillary plane; ACL (anterior cranial base length), from the sella to the foramen caecum; MFL (midfacial length), the minimum distance from the posterior maxillary plane to nasion; LFL (lower facial length), from the anterior nasal spine to the posterior nasal spine; MFP (midfacial projection) from nasion to the foramen caecum (perpendicular to the posterior maxillary plane); FRA (frontal angle) from the metopion to the base of the frontal squama relative to the Frankfurt horizontal; SOL (supraorbital length) from the glabella to fronton (perpendicular to the posterior maxillary plane); GLO (neurocranial curvature or globularity) from the glabella to the opistocranion; and ECV (endocranial volume), which was measured by filling crania with beads; estimates of fossil endocranial volume are from ref. 28. For landmark definitions, see ref. 29.

Received 6 October 1997; accepted 10 February 1998.

1. Day, M. H. & Stringer, C. B. in *L'Homo erectus et la Place de l'Homme de Tautavel parmi les Hominides Fossiles* Vol. 2 (ed. De Lumley, H.) 814–846 (Louis-Jean, Nice, 1982).
2. Lieberman, D. E. Testing hypotheses about recent human evolution from skulls: integrating morphology, function, development and phylogeny. *Curr. Anthropol.* **36,** 159–197 (1995).
3. Kidder, J. H., Jantz, R. L. & Smith, F. H. in *Continuity or Replacement: Controversies in* Homo sapiens *Evolution* (eds Bräuer, G. & Smith, F. H.) 157–177 (A.A. Balkema, Rotterdam, 1992).
4. Lahr, M. M. *The Evolution of Modern Human Cranial Diversity* (Cambridge Univ. Press, 1996).
5. Bilsborough, A. & Wood, B. A. Cranial morphometry of early hominids: facial region. *Am. J. Phys. Antrhropol.* **76,** 61–86 (1988).
6. Enlow, D. H. & Azuma, M. in *Morphogenesis and Malformations of the Face and Brain* (ed. Langman, J.) 217–230 (Harper and Row, New York, 1975).
7. Enlow, D. H. *Facial Growth* 3rd edn (Saunders, Philadelphia, 1990).
8. Bromage, T. G. The ontogeny of *Pan troglodytes* craniofacial architectural relationships and implications for early hominids. *J. Hum. Evol.* **23,** 235–251 (1992).
9. Duterloo, H. S. & Enlow, D. H. A comparative study of cranial growth in *Homo* and *Macaca*. *Am. J. Anat.* **127,** 357–368 (1970).
10. Bromage, T. G. Ontogeny of the early hominid face. *J. Hum. Evol.* **18,** 751–773 (1989).
11. MacCammon, R. *Human Growth and Development* (C. C. Thomas, Springfield, 1970).
12. Weidenreich, F. The brain and its rôle in the phylogenetic transformation of the human skull. *Trans. Am. Phil. Soc.* **31,** 321–442 (1941).
13. Shea, B. T. On aspects of skull form in African apes and orangutans, with implications for hominid evolution. *Am. J. Phys. Anthropol.* **68,** 329–342 (1985).
14. Ravosa, M. J. Ontogenetic perspective on mechanical and nonmechanical models of primate circumorbital morphology. *Am. J. Phys. Anthropol.* **85,** 95–112 (1991).
15. Trinkaus, E. The Neandertal face: evolutionary and functional perspectives on a recent hominid face. *J. Hum. Evol.* **16,** 429–443 (1987).
16. Spencer, M. A. & Demes, B. Biomechanical analysis of masticatory system configuration in Neanderthals and Inuits. *Am. J. Phys. Anthropol.* **91,** 1–20 (1995).
17. Corruccini, R. M. Metrical reconsideration of the Skhul IV and IX and Border Cave 1 crania in the context of modern human origins. *Am. J. Phys. Anthropol.* **87,** 433–445 (1992).
18. Krings, M. *et al*. Neandertal DNA sequences and the origin of modern humans. *Cell* **90,** 19–30 (1997).
19. Stringer, C. B. & Andrews, P. A. Genetic and fossil evidence for the origin of modern humans. *Science* **239,** 1263–1268 (1988).
20. Schwartz, J. H. & Tattersall, I. Significance of some previously unrecognized apomorphies in the nasal region of *Homo neanderthalensis*. *Proc. Natl Acad. Sci. USA* **93,** 10852–10854 (1996).
21. Howells, W. W. *Skull Shapes and the Map* (Peabody Museum Papers no. 79, Cambridge, 1989).
22. Lieberman, D. E. How and why recent humans grow thin skulls: experimental data on systemic cortical robusticity. *Am. J. Phys. Antropol.* **101,** 217–236 (1996).
23. Ruff, C. B., Trinkaus, E. & Holliday, T. W. Body mass and encephalization in Pleistocene *Homo*. *Nature* **387,** 173–176 (1997).
24. Arensburg, B., Schepartz, L. A., Tillier, A. M., Vandermeersch, B. & Rak, Y. A reappraisal of the anatomical basis for speech in Middle Paleolithic hominids. *Am. J. Phys. Anthropol.* **83,** 137–146 (1990).
25. Stevens, K. N. in *Human Communication: A Unified View* (eds David, E. E. & Denes, P. B.) 51–66 (McGraw Hill, New York, 1972).
26. Fant, G. *Acoustic Theory of Speech Production* (Moulton, The Hague, 1960).
27. Lieberman, P. *The Biology and Evolution of Language* (Harvard Univ. Press, Cambridge, MA, 1984).
28. Aiello, L. & Dean, C. *An Introduction to Human Evolutionary Anatomy* (Academic, London, 1990).
29. White, T. D. & Folkens, P. A. *Human Osteology* (Academic, San Diego, 1991).

**Acknowledgements.** I thank B. Arensburg, M. Chech, A. W. Crompton, J.-J. Hublin, L. Humphreey, L. Jellema, R. Kruszinski, P. Langaney, B. Latimer, T. Molleson, D. Pilbeam, A. Poole, C. Ross, G. Sawyer, F. Spoor, C. Stringer, I. Tattersall and A. Walker for help and/or permission to obtain radiographs and CT scans; R. McCarthy, K. Mowbray and R. Bernstein for help with collecting data; and L. Aiello, F. A. Jenkins Jr, P. O'Higgins, O. Pearson, D. Pilbeam, G. P. Rightmire, N. Shubin, R. J. Smith, I. Tattersall and B. Wood for comments on the manuscript. This work was supported by Rutgers University.

Correspondence and requests for materials should be addressed to the author (e-mail: danlieb@rci.rutgers.edu).

# Gene transfer to the nucleus and the evolution of chloroplasts

**William Martin*, Bettina Stoebe†, Vadim Goremykin*, Sabine Hansmann*, Masami Hasegawa‡ & Klaus V. Kowallik†**

* *Institut für Genetik, Technische Universität Braunschweig, Spielmannstrasse 7, 38023 Braunschweig, Germany*
† *Botanisches Institut, Heinrich-Heine-Universität Düsseldorf, Universitätsstrasse 1, 40225 Düsseldorf, Germany*
‡ *The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan*

**Photosynthetic eukaryotes, particularly unicellular forms, possess a fossil record that is either wrought with gaps or difficult to interpret, or both. Attempts to reconstruct their evolution have focused on plastid phylogeny, but were limited by the amount and type of phylogenetic information contained within single genes[1–5]. Among the 210 different protein-coding genes contained in the completely sequenced chloroplast genomes from a glaucocystophyte, a rhodophyte, a diatom, a euglenophyte and five land plants, we have now identified the set of 45 common to each and to a cyanobacterial outgroup genome. Phylogenetic inference with an alignment of 11,039 amino-acid positions per genome indicates that this information is sufficient—but just barely so—to**

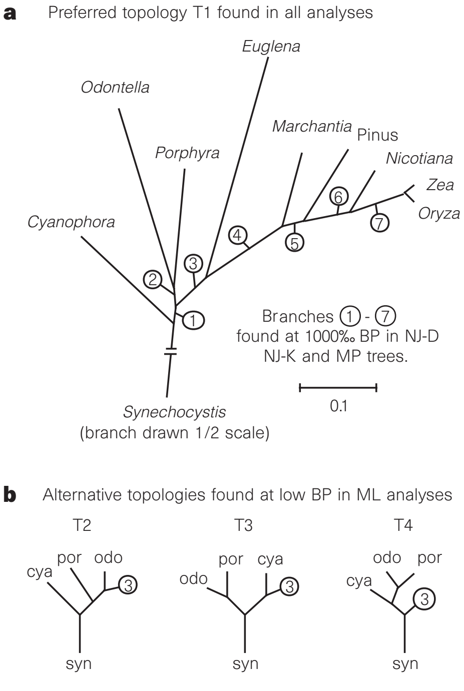
a
Preferred topology T1 found in all analyses
Euglena
Odontella
Marchantia
Pinus
Porphyra
Nicotiana
Zea
Oryza
Cyanophora
1
2
3
4
5
6
7
Branches 1 - 7 found at 1000‰ BP in NJ-D NJ-K and MP trees.
0.1
Synechocystis (branch drawn 1/2 scale)
b
Alternative topologies found at low BP in ML analyses
T2
T3
T4
por
odo
cya
3
syn
por
cya
odo
3
syn
odo
por
cya
3
syn

**Figure 1** Plastid phylogeny interpreted from chloroplast proteins. **a**, Rooted nine species neighbour-joining (NJ) tree of Dayhoff distances for 11,039 amino-acid positions from 45 orthologous proteins common to these chloroplast genomes and *Synechocystis*. All seven branches of this topology (T1) are found in 1,000/1,000 bootstrap samples in maximum parsimony (PROTPARS of PHYLIP) and NJ analysis using either Kimura or Dayhoff distances. The root of the tree is assumed from the model that all plastids sampled here arose from a common chloroplast ancestor. Branches are numbered 1–7 for convenience (see text). The scale bar indicates Dayhoff distance. **b**, Alternative topologies T2, T3 and T4 detected in protein maximum likelihood[10] analyses using the JTT-F model. Taxon abbreviations are given in Methods; branch 3 is the same as in **a**.

**identify the rooted nine-taxon topology. We mapped the process of gene loss from chloroplast genomes across the inferred tree and found that, surprisingly, independent parallel gene losses in multiple lineages outnumber phylogenetically unique losses by more than 4:1. We identified homologues of 44 different plastid-encoded proteins as functional nuclear genes of chloroplast origin, providing evidence for endosymbiotic gene transfer to the nucleus in plants.**

By using literature and database searches, we recorded the presence or absence of the 210 protein-coding genes in nine chloroplast genomes (that is, all known functional protein-coding genes and those designated *ycfs*[6]). We identified 46 protein-coding genes whose sequences are known for all nine genomes. Homologues of these from the *Synechocystis* PCC6803 genome[7] were also retrieved. *Rbc*L was excluded from further analysis because some *rbc*L genes in this taxon sample are related by duplication[8] (paralogy), leaving 45 proteins in the data set. Individual ten-species alignments were constructed for each gene and concatenated to produce a data set consisting of 11,039 amino-acid positions per genome. Trees constructed from these data using the neighbour-joining method[9], with distances estimated using either the Dayhoff matrix or Kimura's method (see Methods) in addition to parsimony analysis, revealed only one topology (T1 in Fig. 1): that is, all seven branches were found in all 1,000 bootstrap samples by all three methods.

T1 is also strongly preferred in protein maximum likelihood (ML)[10] analysis of the concatenated data using the JTT-F model, being found at a frequency of 0.8238 in 10,000 bootstrap samples. But ML detected alternative topologies that the other methods did not: T2 and T3 (Fig. 1) were found with frequencies of 0.0607 and 0.1155, respectively. To investigate alternative topologies further, we used ML to compare the concatenate data analysis to results obtained from individual proteins (Table 1). In individual analyses, only 11 of the 45 proteins prefer T1 over the alternative topologies at $P = 0.95$: *pet*B, *psa*B, *psb*C, *psb*E, *psb*I, *psb*J, *rpl*2, *rpl*14, *rpl*16, *rps*3, *rps*7. T1 is rejected with a difference in log-likelihood larger than twice its standard error by only four proteins: *rps*8 and the subunits of the chloroplast DNA (cpDNA) encoded RNA polymerase *rpo*B, *rpo*C1 and *rpo*C2 (see below). The remaining 30 proteins do not support T1 individually but neither do they reject it at $P = 0.95$. No evidence for widespread lateral gene transfer was detected. Lack of support for T1 by single proteins may be due to sampling error as a result of the small numbers of sites contained within the individual alignments. When the 1,081 sites containing insertions or deletions are excluded from the analysis, bootstrap proportions (BPs) for T1, T3 and T4 undergo only very minor changes (Table 1), indicating that such sites have only a small influence on these data.

The BP for T1 increased from 0.824 (concatenate) to 0.925 across separate analyses (Table 1), raising the question of whether the single tree of concatenated sequences or the combined result of the 45 individual trees is more reliable. To investigate this, we used the Akaike information criterion (AIC)[11], where AIC is defined as $-2\ln L + 2N$, with $\ln L$ being log-likelihood and $N$ the number of parameters. A model that minimizes the AIC is preferable[11]. The JTT-F model uses the amino-acid frequencies of the data (degree of freedom is 19) and, for a tree with ten operational taxonomical units (OTUs), 17 branch lengths must be estimated. Therefore, the number of parameters is 36 ($= 17 + 19$) for each ML analysis. For tree 1 and the concatenated analysis including gap sites in Table 1, AIC was $2 \times 124{,}517.7 + 2 \times 36 = 249{,}107.4$. For T1 from separate analyses, AIC was $2 \times 118{,}401.8 + 2 \times 36 \times 45 = 240{,}043.6$. Thus, AIC for the separate analyses is lower than that for the concatenated analysis. This indicates that the separate analyses approximate the underlying evolutionary process better than the concatenated analysis, which assumes homogeneity of substitution process across genes. This, in turn, indicates that the combined result of separate analyses should be more reliable than the one obtained from the concatenated data. Therefore, the higher BP of T1 obtained across separate analyses compared with concatenated analysis provides further support for T1. All three subunits of the cpDNA-encoded RNA polymerase (*rpo*B, *rpo*C1 and *rpo*C2) strongly reject T1 and support T2, indicating that *Odontella*'s *rpo* genes tend to branch with their homologues from the chlorophyte lineage. This could reflect paralogy, stochastic or other[12] factors that may have differentially affected the evolution of *rpo* genes in the *Odontella*/*Porphyra* lineages. However, when the *rpo* genes are excluded from the analysis, the BP for T1 does not increase; rather, it decreases to 0.72 (Table 1), probably because the *rpo* genes reject T3 and T4 even more strongly than T1.

We observed differences in amino-acid composition across genomes: A + T-rich genomes (that is, *Euglena*) tended to possess higher frequencies of A + T-rich codons, whereas the converse was observed for *Synechocystis* (G+C-rich). This might influence the substitution process at sites under low constraint. We therefore used a variant of ML taking into account the suggestion[13] that amino acids which frequently substitute may influence topology. We carried out ML analysis after converting all valine residues into isoleucines and all lysine residues into arginines ($V = I$, $K = R$) or including leucine and methionine into the first group ($I = V = L = M$, $R = K$), but did not obtain significant differences from the analysis of the original data in either case.

Therefore, ML strongly favours T1 in all types of analysis. Furthermore, support for T1 increases as the amount of data analysed increases (Table 1). Although ML, like all methods of phylogenetic inference, can fail if the assumed substitution model is incorrect[14–16], the other methods do not detect topologies other than T1 at all. The most straightforward interpretation of these findings is that chloroplast genomes (just barely) contain enough information to infer fully the properly rooted phylogeny of these nine photosynthetic organelles.

Chloroplast proteins resoundingly support branches 4–7 in Fig. 1—branches that are known to be correct from the fossil record—

**Table 1 Topologies of cpDNA-encoded proteins detected by maximum likelihood**

| No. of genes | Type of sites | No. of sites | Type of analysis | T1 | | T2 | | T3 | | T4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ln*L* | BP* | Δln*L* ± s.e. | BP | Δln*L* ± s.e. | BP | Δln*L* ± s.e. | BP |
| 45 | All | 11,039 | Separate | −118,401 | 9,246 | 177.9 ± 48.4† | 1 | 47.1 ± 43.4 | 647 | 65.2 ± 41.7 | 106 |
| | | | Concatenate | −124,517 | 8,238 | 64.1 ± 39.7 | 607 | 36.9 ± 35.0 | 1,155 | 99.2 ± 25.6† | 0 |
| | Excl. gaps | 9,958 | Separate | −103,207 | 9,030 | 194.0 ± 48.2† | 0 | 41.0 ± 42.3 | 776 | 54.2 ± 41.0 | 194 |
| | | | Concatenate | −107,972 | 8,367 | 109.8 ± 38.7† | 21 | 29.4 ± 28.6 | 1,606 | 71.5 ± 25.1† | 6 |
| 42‡ | All | 7,864 | Separate | −72,762 | 7,199 | 247.4 ± 48.4† | 0 | 32.2 ± 35.4 | 730 | 19.7 ± 35.5 | 2,071 |
| | | | Concatenate | −76,328 | 6,880 | 232.7 ± 36.81† | 0 | 23.7 ± 23.1 | 1,623 | 23.2 ± 23.0 | 1,497 |
| | Excl. gaps | 7,660 | Separate | −71,010 | 7,670 | 252.0 ± 48.1† | 0 | 33.3 ± 35.6 | 736 | 24.2 ± 35.6 | 1,594 |
| | | | Concatenate | −74,531 | 7,451 | 236.3 ± 36.8† | 0 | 22.8 ± 23.4 | 1,365 | 26.1 ± 23.0 | 1,184 |

Δln*L* ± s.e. indicates the difference in log likelihood relative to T1 (the preferred topology in each case) and one standard error of that difference. The TOTALML program in MOLPHY[10] was used to evaluate the total evidence of separate analyses of individual proteins.

* Bootstrap proportion for topology in 10,000 replicates × $10^4$ using the RELL method[10].

† Significantly worse than T1 ($\Delta\ln L > 2$s.e.).

‡ Excluding *rpoB*, *rpoC1* and *rpoC2*, which favour T2 and reject T1 at $P = 0.95$ (see text).

and clarify several unresolved issues in the early evolution of photosynthetic organelles. First, branch 3 confirms the chlorophytic ancestry of *Eugelena*'s three-membrane bound plastids, as first suggested by Gibbs[17], who argued that *Euglena*'s plastids arose through secondary symbiosis by capture of a eukaryotic chlorophyte in the kinetoplastid lineage (trypanosomes and relatives), with loss of the chlorophyte nucleus. Common ancestry of the euglenoid and trypanosomatid nucleocytoplasmic lineages has never been seriously challenged[18], and is borne out by 18S ribosomal RNA and protein data[19]. *Euglena*'s plastid ancestry, however, has been unrecoverable with rRNA data. Most analyses grouped its plastids with rhodophytes and only a few detected branching with chlorophytes, with minimal statistical support[3,4,15]. In the case of rRNA, this is partly due to base-composition bias[15], but is more generally attributable to the limited information contained in any individual gene. Our data also strongly support a common ancestry for the diatom (*Odontella*) and rhodophyte (*Porphyra*) plastid genomes. In the light of molecular and cytological evidence indicating a common descent of the diatom nuclear lineage with oomycetes[20], our findings reflect a secondary symbiotic origin of *Odontella*'s four membrane-bound plastids[21]. The chlorophyll *a* and phycobilin-containing cyanelles of the glaucocystophyte *Cyanophora paradoxa*, which have retained not only phycobilisomes but also a eubacterial peptidoglycan wall[3,22], branch basal to the chlorophyte and rhodophyte plastids, a position that has been impossible to establish with confidence using data from single genes[3,4]. The data analysed here provide strong support for this basal position.

We have mapped onto the topology T1 of Fig. 1 all 205 genes demonstrated to exist in the cpDNA molecule ancestral to the genomes surveyed (Fig. 2). Gene-loss events that are unique to individual lineages account for the fate of only 58 genes across the tree. The majority of genes (101) have undergone parallel loss in independent lineages. In a survey of only nine genomes, we found that 44 genes were lost twice independently, 43 genes have undergone three parallel losses, and 14 genes have been lost four times in independent lineages. The numbers of unique (6) and parallel (40) events indicated in Fig. 2 reflect the unlikely but event-minimizing premise that loss occurred in the form of the indicated gene blocks. The visible, ongoing process of *ndh* gene loss from *Pinus* cpDNA[23]

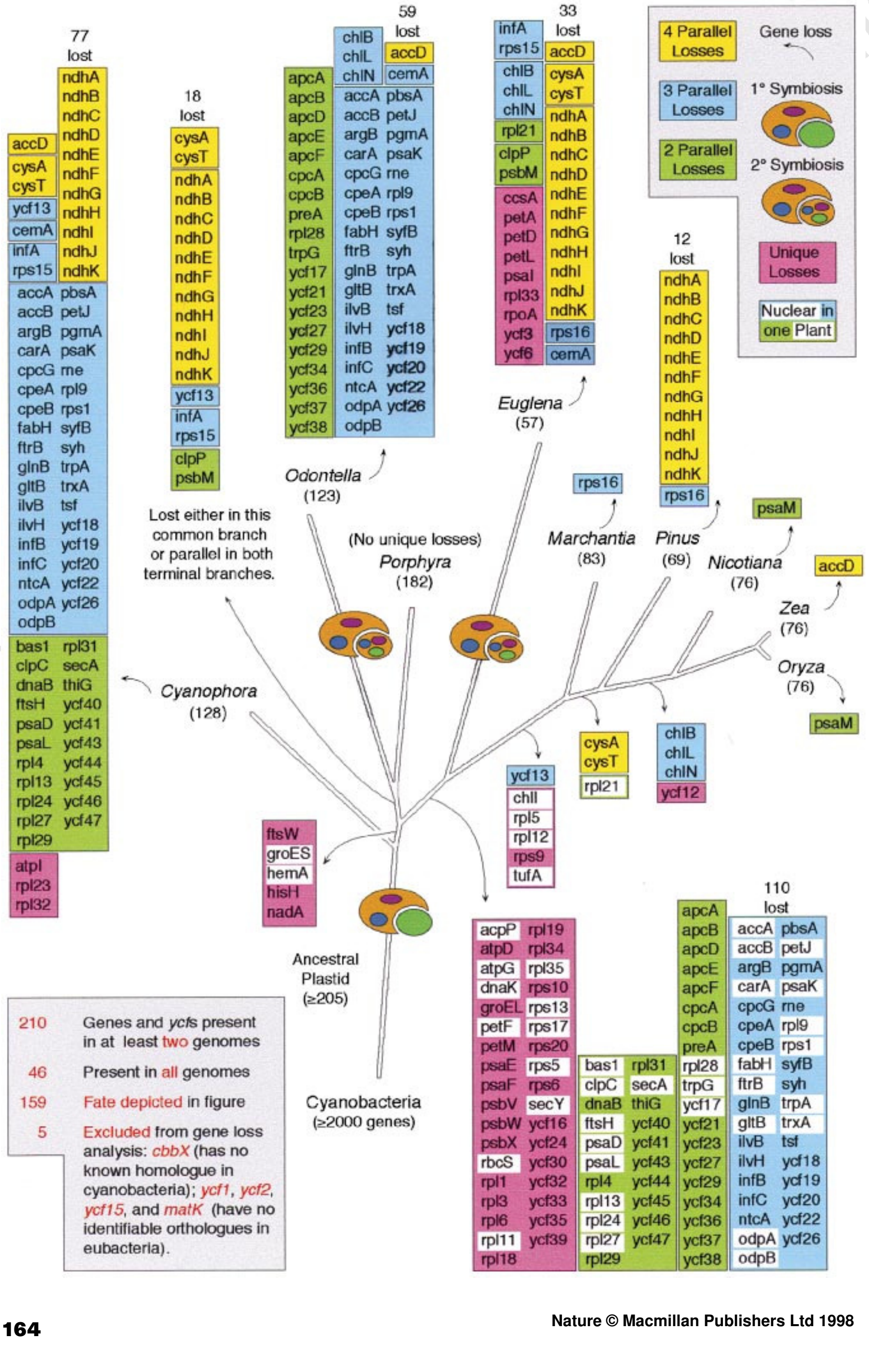
77 lost
ndhA ndhB ndhC ndhD ndhE ndhF ndhG ndhH ndhI ndhJ ndhK
accD cysA cysT ycf13 cemA infA rps15
accA pbsA accB petJ argB pgmA carA psaK cpcG rne cpeA rpl9 cpeB rps1 fabH syfB ftrB syh glnB trpA gltB trxA ilvB tsf ilvH ycf18 infB ycf19 infC ycf20 ntcA ycf22 odpA ycf26 odpB
bas1 rpl31 clpC secA dnaB thiG ftsH ycf40 psaD ycf41 psaL ycf43 rpl4 ycf44 rpl13 ycf45 rpl24 ycf46 rpl27 ycf47 rpl29
atpI rpl23 rpl32
Cyanophora
(128)
18 lost
cysA cysT ndhA ndhB ndhC ndhD ndhE ndhF ndhG ndhH ndhI ndhJ ndhK ycf13 infA rps15 clpP psbM
Lost either in this common branch or parallel in both terminal branches.
59 lost
chlB chlL chlN accD cemA
apcA apcB apcD apcE apcF cpcA cpcB preA rpl28 trpG ycf17 ycf21 ycf23 ycf27 ycf29 ycf34 ycf36 ycf37 ycf38
accA pbsA accB petJ argB pgmA carA psaK cpcG rne cpeA rpl9 cpeB rps1 fabH syfB ftrB syh glnB trpA gltB trxA ilvB tsf ilvH ycf18 infB ycf19 infC ycf20 ntcA ycf22 odpA ycf26 odpB
Odontella
(123)
(No unique losses)
Porphyra
(182)
33 lost
infA rps15 accD chlB chlL chlN cysA cysT rpl21 ndhA ndhB clpP ndhC psbM ndhD
ccsA ndhE petA ndhF petD ndhG petL ndhH psaI ndhI rpl33 ndhJ rpoA ndhK ycf3 rps16 ycf6 cemA
Euglena
(57)
rps16
Marchantia
(83)
12 lost
ndhA ndhB ndhC ndhD ndhE ndhF ndhG ndhH ndhI ndhJ ndhK rps16
Pinus
(69)
psaM
Nicotiana
(76)
accD
Zea
(76)
Oryza
(76)
psaM
4 Parallel Losses
3 Parallel Losses
2 Parallel Losses
Unique Losses
Nuclear in one Plant
Gene loss
1° Symbiosis
2° Symbiosis
cysA cysT rpl21
chlB chlL chlN ycf12
ycf13 chlI rpl5 rpl12 rps9 tufA
ftsW groES hemA hisH nadA
Ancestral Plastid
(≥205)
Cyanobacteria
(≥2000 genes)
110 lost
acpP rpl19 atpD rpl34 atpG rpl35 dnaK rps10 groEL rps13 petF rps17 petM rps20 psaE rps5 psaF rps6 psbV secY psbW ycf16 psbX ycf24 rbcS ycf30 rpl1 ycf32 rpl3 ycf33 rpl6 ycf35 rpl11 ycf39 rpl18
bas1 rpl31 clpC secA dnaB thiG ftsH ycf40 psaD ycf41 psaL ycf43 rpl4 ycf44 rpl13 ycf45 rpl24 ycf46 rpl27 ycf47 rpl29
apcA apcB apcD apcE apcF cpcA cpcB preA rpl28 trpG ycf17 ycf21 ycf23 ycf27 ycf29 ycf34 ycf36 ycf37 ycf38
accA pbsA accB petJ argB pgmA carA psaK cpcG rne cpeA rpl9 cpeB rps1 fabH syfB ftrB syh glnB trpA gltB trxA ilvB tsf ilvH ycf18 infB ycf19 infC ycf20 ntcA ycf22 odpA ycf26 odpB
210 Genes and ycfs present in at least two genomes
46 Present in all genomes
159 Fate depicted in figure
5 Excluded from gene loss analysis: cbbX (has no known homologue in cyanobacteria); ycf1, ycf2, ycf15, and matK (have no identifiable orthologues in eubacteria).

**Figure 2** Phylogenetic distribution of gene loss from chloroplast genomes. Colour keys designating frequency of parallel gene losses are given at top right. Numbers below species names indicate the number of protein coding genes and *ycf*s in the corresponding chloroplast genome. Numbers above gene columns represent the number of genes lost which are accounted for in the figure for the given genome. The symbols for primary and secondary symbiosis are indicated. Five genes were excluded from gene-loss analysis for reasons indicated at the lower left. Some highly divergent proteins may have escaped detection with BLAST searches. Functional, transferred nuclear homologues of chloroplast origin are indicted in white rectangles. In *Pinus*, four *ndh* genes are completely missing (*ndhA*, *ndhF*, *ndhG*, *ndhJ*), the other seven are pseudogenes[23] and are scored as losses here.

indicates that genes are lost individually, rather than as blocks; the true numbers of parallel events are therefore probably much greater. If we use a more realistic model and count each loss of an individual gene as one event, the ratio of parallel to unique events becomes 4.7 (58 unique versus 273 parallel), an overwhelming excess of parallel losses.

Some of these genes have been lost altogether, for example constituents of phycobilisomes (*apc*, *cpc* and *cpe* gene products) in the chlorophyte and *Odontella* lineages, whereas others have been transferred to the nucleus. We have documented 44 bona fide cases of functional plant nuclear genes among the 210 genes examined that descend from cyanobacterial genomes (white rectangles in Fig. 2). These genes became expressed by the eukaryotic transcription machinery and acquired transit peptides for reimport of the encoded products into the organelle of their genetic origin[8,24].

Why should plastids tend to donate genes to the nucleus at all? Nuclear primacy in gene regulation has many advantages[24], but population genetic factors may also be important. Notably, the effects of Muller's ratchet—the rapid accumulation of deleterious mutations in asexual populations—can be observed in metazoan mitochondrial genomes[25] and in endosymbiotic bacteria[26]. Transfer of a gene from cpDNA to the nucleus should increase its rate of recombination and reduce its genetic load. This would generally favour nuclear fixation and could help to account for the degree of successful gene transfer indicated in Fig. 2. Chloroplast genomes furnish plant phylogenetics with a strong backbone and provide a simple model to study fragmented prokaryotic genomes dispersed across eukaryotic chromosomes[8,27]. □

## Methods

Complete sequences of *Zea mays* (X86563; zea), *Oryza sativa* (X15901; ory), *Nicotiana tabacum* (S54304; nic), *Pinus thunbergii* (D17510; pin), *Marchantia polymorpha* (X04465; mar), *Euglena gracilis* (X70810; eug), *Porphyra purpurea* (U38804; por), *Odontella sinensis* (Z67753; odo), and *Cyanophora paradoxa* (U30821; cya) chloroplast genomes were retrieved from GenBank. Starting with the *Porphyra* sequence, BLAST searches were made with all designated encoded proteins and annotated open reading frames. Homologues from the other eight cpDNAs and from *Synechocystis* PCC6803 (ref. 7) (syn) were retrieved. For genes missing in at least one chloroplast genome by this search, local TFASTA searches were made against six-frame translations of the nucleotide sequences as a control. The set of 45 orthologous proteins encoded in all 10 genomes were aligned in individual files with PILEUP of the Genetics Computer Group (GCG) package, and written into PHYLIP[28] format with CLUSTALW[29].

The alignment of ten OTUs with 11,039 positions in each sequence was used as input for Fig. 1. For distance and parsimony analysis, the concatenated data was bootstrapped[28]; for ML analysis, the RELL bootstrap method[10] was used. Using the 11,039-site concatenated data, we calculated the approximate ln*L* (ref. 10) for all 2,027,025 possible trees, and the best 1,000 trees by the approximate likelihood criterion were provided for full likelihood analysis. T1 had the highest likelihood. All trees that disrupted any of branches 3–7 were significantly ($\Delta \ln L > 5SE$) worse than T1. Therefore, branches 3–7 were defined, and the remaining 15 possible trees were more extensively examined using various subsets of the complete data (Table 1). Names of the 45 protein coding genes used for phylogenetic inference are given alphabetically, lengths of alignment used are in parentheses: *atpA* (494); *atpB* (491); *atpE* (237); *atpF* (182); *atpH* (81); *petB* (215); *petG* (38); *psaA* (754); *psaB* (738); *psaC* (82); *psaJ* (38); *psbA* (343); *psbB* (507); *psbC* (460); *psbD* (353); *psbE* (72); *psbF* (40); *psbH* (58); *psbI* (36); *psbJ* (42); *psbK* (46); *psbL* (39); *psbN* (43); *psbT* (31); *rpl2* (277); *rpl14* (124); *rpl16* (134); *rpl20* (115); *rpl22* (111); *rpl36* (38); *rpoB* (1098); *rpoC1* (689); *rpoC2* (1388); *rps2* (225); *rps3* (250); *rps4* (211); *rps7* (156); *rps8* (144); *rps11* (146); *rps12* (123); *rps14* (103); *rps18* (60); *rps19* (94); *ycf4* (175); *ycf9* (58).

In maize, it is estimated that about 25 sites in cpDNA are subject to RNA editing[30]. Although the total degree of editing is not known for all OTUs considered here, it will only affect comparison of differentially edited codons at the small fraction of edited sites, and was therefore neglected. The concatenated alignment contains 4,180 constant and 6,859 polymorphic sites. Criteria for indicating the presence of a functional nuclear gene (Fig. 2) were (1) expression (full-size, or nearly so, cDNA sequence characterized); (2) presence of a functional or putative transit peptide; and (3) greater similarity between cpDNA and nuclear homologues than between cyanobacterial and nuclear homologues. BLAST results apply to GenBank on 15 October 1997. The 32 amino-acid fragment of *ycf1* present in rice cpDNA (JQ0282) was neglected. To avoid confusion in Fig. 2, red and green[8] type *rbcL* and *rbcS* are counted and mapped as two genes, rather than four. *PetL*, a 31-amino-acid long component of the cytochrome $b_6/f$ complex, is too short to detect its *Synechocystis* homologue (gi1653694) among the highest BLAST scores, but was counted as being of cyanobacterial origin; the same applies for *ycf33* (67 amino acids long). BLAST search results, ML results for individual proteins, accession numbers for identified nuclear homologues, concatenate and individual alignments used in this study can be retrieved from 134.169.70.80/ftp/pub/incoming/.

Received 26 August 1997; accepted 4 February 1998.

1. Palenik, B. & Haselkorn, R. Multiple evolutionary origins of prochlorophytes, the chlorophyll *b*-containing prokaryotes. *Nature* **355,** 265–267 (1992).
2. Urbach, E., Robertson, D. L. & Chisolm, S. W. Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. *Nature* **355,** 267–270 (1992).
3. Helmchen, T. A., Bhattacharya, D. & Melkonian, M. Analyses of ribosomal RNA sequences from glaucocystophyte cyanelles provide new insights into the evolutionary relationships of plastids. *J. Mol. Evol.* **41,** 203–210 (1995).
4. Van de Peer, Y., Rensing, S., Maier, U.-G & De Wachter, R. Substitution rate calibraiton of small subunit RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. *Proc. Natl Acad. Sci. USA* **93,** 7744–7748 (1996).
5. Melkonian, M. Systematics and evolution of the algae: Endocytobiosis and the evolution of the major algal lineages. *Progr. Bot.* **57,** 281–311 (1996).
6. Hallick, R. B. *et al*. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.* **21,** 3537–3544 (1993).
7. Kaneko, T. *et al*. Sequence analysis of the gneome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assigment of potential protein-coding regions. *DNA Res.* **3,** 109–136 (1996).
8. Martin, W. & Schnarrenberger, C. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr. Genet.* **32,** 1–18 (1997).
9. Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
10. Adachi, J. & Hasegawa, M. Computer Science Monographs, No. 28. MOLPHY Version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood. (Institute of Statistical Mathematics, Tokyo, 1996).
11. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* CA**19,** 716–723 (1974).
12. Hedtke, B., Börner, T. & Weihe, A. Mitochondrial and chloroplast phage-like RNA polymerases in *Arabidopsis*. *Science* **277,** 809–811 (1997).
13. Naylor, G. J. P. & Brown, W. M. Structural biology and phylogenetic estimation. *Nature* **388,** 527–528 (1997).
14. Adachi, J. & Hasegawa, M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42,** 459–468 (1996).
15. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11,** 605–612 (1994).
16. Lockhart, P. J., Larkum, A. W., Steel, M., Waddell, P. J. & Penny, D. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl Acad. Sci. USA* **93,** 1930–1934 (1996).
17. Gibbs, S. P. The chloroplast of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot.* **56,** 2883–2889 (1978).
18. Cavalier-Smith, T. Kingdom Protozoa and its 18 phyla. *Microbiol. Rev.* **57,** 953–994 (1993).
19. Henze, K. *et al*. A nuclear gene of eubacterial origin in *Euglena* reflects cryptic endosymbioses during protist evolution. *Proc. Natl Acad. Sci. USA* **92,** 9122–9126 (1995).
20. Cavalier-Smith, T. *et al*. Cryptomonad nuclear and nucleomorph 18S rRNA phylogeny. *Eur. J. Phycol.* **31,** 315–328 (1996).
21. Kowallik, K. V., Stöbe, B., Schaffran, I., Kroth-Pancic, P. & Freier, U. The chloroplast genome of a chlorophyll *a+c* containing alga, *Odontella sinensis*. *Plant Mol. Biol. Reptr.* **13,** 336–342 (1995).
22. Löffelhardt, W. & Bohnert, H. J. in *The Molecular Biology of Cyanobacteria* (ed. Bryant, D. A.) 56–89 (Kluwer, Dordrecht, 1994).
23. Wakasugi, T. *et al*. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of black pine *Pinus thunbergii*. *Proc. Natl Acad. Sci. USA* **91,** 9794–9798 (1994).
24. Herrmann, R. G. in *Eukaryotism and Symbiosis* (eds Schenk, H. E. A., Herrmann, R. G., Jeon, K. W. & Schwemmler, W.) 73–118 (Springer, Heidelberg, 1997).
25. Lynch, M. Mutation accumulation in transfer RNAs: Molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol. Biol. Evol.* **13,** 209–220 (1996).
26. Moran, N. A. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93,** 2873–2878 (1996).
27. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392,** 37–41 (1998).
28. Felsenstein, J. PHYLIP (Phylogeny inference package) manual, version 3.5c (Univ. Washington, Seattle, Dept Genetics, 1993).
29. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
30. Maier, R. M., Neckermann, K., Igloi, G. L. & Kössel, H. Complete sequence of maize chloroplast genome: Gene content, hotspots of divergence and the tuning of genetic information by transcript editing. *J. Mol. Biol.* **251,** 614–628 (1995).

**Acknowledgements.** We thank R. G. Herrmann for discussion, H. Phillipe for comments on the manuscript, the Rechenzentrum der Universität Braunschweig for use of computer facilities, and the Deutsche Forschungsgemeinschaft (W.M. and K.V.K.) and the Ministry of Education, Science, Sports and Culture of Japan (M.H.) for financial support. B.S. is the recipient of a psotdoctoral stipend from the DFG; V.G. is the recipient of a Ph.D. stipend from DAAD.

Correspondence and requests for materials should be addressed to W.M. (e-mail: w.martin@tu-bs.de).