BMC Evolutionary Biology

Adaptive Evolution of Chloroplast Genome Structure Inferred Using a Parametric Bootstrap Approach

BMC Evolutionary Biology 2006, 6:13 doi:10.1186/1471-2148-6-13

Liying Cui (liying@psu.edu)
Jim Leebens-Mack (jhl10@psu.edu)
Li-San Wang (lswang@med.upenn.edu)
Jijun Tang (jtang@cse.sc.edu)
Linda Rymarquis (lar24@cornell.edu)
David B. Stern (ds28@cornell.edu)
Claude W. dePamphilis (cwd3@psu.edu)

ISSN 1471-2148

Article type Research article

Submission date 15 Jun 2005

Acceptance date 9 Feb 2006

Publication date 9 Feb 2006

Article URL http://www.biomedcentral.com/1471-2148/6/13

Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach

Liying Cui[1], Jim Leebens-Mack[1], Li-San Wang[2], Jijun Tang[3], Linda Rymarquis[4], David B. Stern[4] and Claude W. dePamphilis[1§]

[1]Department of Biology, Institute of Molecular Evolutionary Genetics, and Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA
[2]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
[4]Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA

[§]Corresponding author

Email addresses:
LC: liying@psu.edu
JLM: jhl10@psu.edu
LSW: lswang@med.upenn.edu
JT: jtang@cse.sc.edu
LR: lar24@cornell.edu
DBS: ds28@cornell.edu
CWD: cwd3@psu.edu

Abstract

Background

Genome rearrangements influence gene order and configuration of gene clusters in all genomes. Most land plant chloroplast DNAs (cpDNAs) share a highly conserved gene content and with notable exceptions, a largely co-linear gene order. Conserved gene orders may reflect a slow intrinsic rate of neutral chromosomal rearrangements, or selective constraint. It is unknown to what extent observed changes in gene order are random or adaptive. We investigate the influence of natural selection on gene order in association with increased rate of chromosomal rearrangement. We use a novel parametric bootstrap approach to test if directional selection is responsible for the clustering of functionally related genes observed in the highly rearranged chloroplast genome of the unicellular green alga *Chlamydomonas reinhardtii*, relative to ancestral chloroplast genomes.

Results

Ancestral gene orders were inferred and then subjected to simulated rearrangement events under the random breakage model with varying ratios of inversions and transpositions. We found that adjacent chloroplast genes in *C. reinhardtii* were located on the same strand much more frequently than in simulated genomes that were generated under a random rearrangement processes (increased sidedness; p <0.0001). In addition, functionally related genes were found to be more clustered than those evolved under random rearrangements (p < 0.0001). We report evidence of co-transcription of neighboring genes, which may be responsible for the observed gene clusters in *C. reinhardtii* cpDNA.

Conclusions

Simulations and experimental evidence suggest that both selective maintenance and directional selection for gene clusters are determinants of chloroplast gene order.

Background

The influence of genotype on phenotype is not limited to the coding of peptides and functional RNAs by nucleotide sequences. An organism's phenotype is also affected by the chromosomal arrangement of genes and the interaction of gene products. Comparative genomics has revealed a variety of gene clusters and chromosomal segments that have remained intact over hundreds of millions of years [1]. Selection for clustering of co-transcribed genes has been hypothesized to influence gene order within bacterial and organelle genomes, where gene clusters typically encode multiple components of a functional pathway [2]. For example, the ribosomal proteins are encoded by similar operons in archaebacteria, eubacteria and plastids [3]. In eukaryotic genomes, co-expression of neighboring genes is significantly associated with the functional roles of the genes (such as housekeeping genes or genes in the same metabolic pathway) [4, 5]. One way that those genes become clustered is through tandem duplication, which usually results in functionally related genes being adjacent. On the other hand, unrelated genes may also be brought together through chromosome rearrangements (recombination, inversion and transposition).

Unless selection is acting to maintain or promote gene clusters, gene orders in genomes subjected to rearrangements should become randomized with respect to function or co-expression profiles. Significant clustering has been inferred using permutation tests that compare observed physical distances between pairs or blocks of co-expressed or functionally related genes to a null distribution constructed from randomized gene orders

[4, 5]. However, this approach is limited since the evolutionary history of the genome was not considered. When comparing gene orders among related species, it is possible to estimate the ancestral genome and to simulate a null distribution for changes in gene order using a model. This evolutionary approach can be used to test directly the influence of selection on genome structure, that is, whether present-day genome structure has been influenced by directional selection for clustering of functionally related genes.

Small genomes, especially those of organelles and bacteria, are well suited to global comparisons of gene order. Like eukaryotic genomes, they are subject to structural changes such as inversion, transposition or translocation, as well as gene loss and (more rarely) gene gain. Chloroplast DNAs in most land plants share a highly conserved gene content and similar gene orders [6]. Most cpDNAs include two identical regions in opposite orientations called the inverted repeat (IR), flanked by large single copy (LSC) and small single copy (SSC) regions. The IRs generally contain the bacterial-like rRNA gene clusters, and the genes involved in photosynthesis (photosystem I/II, cytochrome b_6/f, and ATP synthase) are arranged similarly in chloroplast and cyanobacterial genomes [2, 3, 7]. Despite these well-characterized patterns, it is unknown to what extent the conserved gene order reflects a slow intrinsic rate of neutral chromosomal rearrangements, rather than selection against alternative gene orders. A model of neutral rearrangement of gene order is required to test formally whether gene orders evolve under selection which prefers some gene arrangements over others.

Nadeau and Taylor first proposed a model for the neutral evolution of gene order in comparisons of mouse and human chromosomes [8]. This "random breakage model" provides a null hypothesis for the evolution of gene order. It assumes a random distribution of break points and allows all possible gene orders without restrictions. The random breakage model has been used to infer organismal phylogenies from gene order data [9]. The gene order difference can be measured using the inversion distance, which is the minimal number of inversions necessary to transform one gene order to another. Currently, the most accurate heuristic approach is implemented in the GRAPPA software [10], which is generally suitable for small taxon sets because the algorithm scores inversion medians for all nodes iteratively across all possible phylogenies. Algorithms for genomes with arbitrary rearrangements, a few deletions and duplications have been developed [11], and the capacity of GRAPPA can be scaled up with the discovering method (DCM) to potentially very large data sets [12].

The random breakage model does not account for recombination hotspots, which have been reported from human-mouse genome comparisons [13]. However, at this time it may be difficult to model these hotspots, because the precise locations of reused breakpoints are unknown due to insufficient resolution of gene orders and potential errors in homology assessment given the scale of eukaryotic chromosomes [14]. Thus, the fragile breakage model [13], as an alternative to the random breakage model, has not been well established.

Whereas gene order is generally conserved among land plant cpDNAs, very little synteny is observed between this group and cpDNAs of the chlorophytic green algae *C. reinhardtii* [15, 16] and *Chlorella vulgaris* [17]. The apparently increased rearrangement rate is associated with invasion by a large number of short dispersed repeats (SDRs), for which the evolutionary distribution is still poorly defined. The large number of rearrangements provides an excellent opportunity to test whether natural selection has preferred some changes in gene order. Here we present novel statistics and parametric tests that lead us to reject the models of random rearrangement in favor of directional selection for clustering of functionally related genes in *C. reinhardtii*. We also present experimental evidence that adaptive evolution of chloroplast genome structure could be driven by the advantage of concerted regulation conferred by polycistronic transcription.

Results

Functional clusters are not randomly distributed

We compared gene orders of representative cpDNAs from land plants, including tobacco (*Nicotiana tabacum*, [GenBank:NC_001879]) [18] and liverwort (*Marchantia polymorpha*, [GenBank:NC_001319]) [19], a charophytic green alga (*Chaetosphaeridium globosum* [GenBank:NC_004115]) [20], chlorophytic green algae (*Nephroselmis olivacea* [GenBank:NC_000927] [21], *C. vulgaris* [GenBank:NC_001865] [17], *C. reinhardtii* [GenBank:BK000554] [16]), a green flagellate alga with uncertain affinities (*Mesostigma viride* [GenBank:NC_002186]) [22], and the plastid of *Cyanophora paradoxa* [GenBank:NC_001675] [23] (Figure 1) (Additional file 1, Additional file 3). To measure the genome structure in terms of clustering by chromosome locations and gene functions, we defined "sided blocks" as contiguous genes coded on the same strand of the plastid chromosome, and "functional clusters" as blocks of functionally related genes (see Methods). The randomness in the observed distribution of shared genes in chloroplast genomes with respect to gene function was assessed using a Kolmogorov-Smirnov test. The null hypothesis was rejected in all seven cpDNAs investigated for genes in functional categories such as ATP synthases and electron transport ($p<<0.05$, Table 1). While this test suggests some degree of functional clustering in all chloroplast genomes, it does not take into account the phylogenetic relationship of these organisms, so it is unclear whether functional clustering in chloroplast genomes is a legacy of genome organization in a cyanobacteria-like ancestor, or the product of selection on gene order in the face of genome rearrangements.

Extensive rearrangements from the ancestral chloroplast genome to *C. reinhardtii*

In order to investigate evolutionary changes of gene order, we constructed a phylogeny of seven representative cpDNAs and rooted with the sequence of *C. paradoxa* [23]. Maximum parsimony, neighbor joining and maximum likelihood analyses of an alignment of 50 concatenated protein sequences including a total of 19,836 aligned sites (Additional file 2), all yielded identical fully resolved topologies with high bootstrap support (Figure 2A). *Mesostigma* was placed as a basal charophyte lineage in one previous analysis [24]. The unrooted phylogeny of seven cpDNAs (Figure 2B) is congruent with the alternative placement of *Mesostigma* either as a basal charophyte [24] or basal to both charophyte and chlorophyte lineages [22]. This tree was used as the reference phylogeny for gene order inference.

We scored the orders of 85 genes shared in the seven genomes (Gene orders are in additional file 3). Then we used modified versions of GRAPPA [11, 25] to compute the inversion distance between ancestral nodes and each terminal node (Figure 2B; see Methods). The branches leading to two chlorophytic green algae, *C. reinhardtii* and *C. vulgaris*, are much longer than the branches leading to the other taxa, and that many more steps were inferred on the *C. reinhardtii* lineage relative to the *C. vulgaris* lineage. Gene duplications or deletions were mapped before scoring the ancestral genomes with inversions, and were not counted as rearrangements. IRs were present in all inferred ancestral nodes, and one copy was lost in *C. vulgaris*. Ancestral gene orders were scored on all the phylogenies using a two-step approach (see Methods). Due to the computational time limit (the full search for ancestral gene orders may require months), we stopped scoring all possible ancestral gene orders with the data set after 25 days and took the best scored ancestral gene orders at that time (Additional file 4).

The cpDNAs of two land plants, *N. tabacum* and *M. polymorpha*, were separated by an estimated 7 inversions based on the data set. One large inversion (~ 30 kb) in the LSC region has long been recognized to separate the two genomes [26]. Additional rearrangements are directly observable through comparison of gene order files for the two species (see additional file 5 for the sequences of gene order rearrangements). Using GRAPPA, all rearrangements were inferred as inversions, but the total number of inversion events estimated by GRAPPA may be greater than the true (but unknown) mixture of inversions and transpositions because one transposition could result in the same change in gene order as two or three inversions.

Increased order in the genome structure after rearrangements

Two genomic structural characteristics were measured: the propensity of adjacent genes to be clustered on the same strand (using the sidedness index C_s) and the clustering of functionally related genes (using the functional cluster index, C_f) (see Methods). Both indices were calculated for the inferred ancestral gene orders and extant daughter lineages. Among land plants and charophytes, the inferred sidedness among ancestral genomes was similar to extant lineages, however, among the chlorophytes an opposite trend was observed, especially in the *C. reinhardtii* lineage (Additional file 3). The large number of rearrangements in the *C. reinhardtii* cpDNA lineage resulted in dramatically increased sidedness relative to the inferred most recent common ancestor of *C.*

reinhardtii and *C. vulgaris* (C_s ancestor = 0.6966, C_s observed = 0.8710; Figure 3A). A small increase of C_s was found in the *N. olivacea* lineage and there was almost no change in the lineage leading to *C. vulgaris*. A large increase was also observed in the functional clustering index, C_f, for *C. reinhardti* (C_f ancestor= 0.01674, C_f observed = 0.03397; Figure 3B), whereas the trend was less profound in other lineages (Additional file 3). Thus, even if the ancestral genome already had a "sided" structure, sidedness increased with genome rearrangements as *C. reinhardtii* cpDNA evolved. The inferred increase in sidedness and functional clustering in the face of the large number of rearrangements on the lineage leading to *C. reinhardtii* might be adaptive, if such increases were not expected under random rearrangements.

To test the null hypothesis that the changes in C_s and C_f were consequence of random genome rearrangements rather than a consequence of directional selection (H_0: random rearrangement; H_A: constraints in rearrangements), we simulated random rearrangements starting with the inferred ancestral genome along the branch leading to *C. reinhardtii*. Although inversions are the most abundant type of rearrangement in cpDNAs [27], we also considered the contribution of transpositions under three inversion to transposition ratios, while the total number of rearrangements was fixed according to the branch length inferred using GRAPPA (Figure 2B). Three simulations with 10,000 replicates were conducted with inversion to transposition ratios of 1:0, 10:1 and 1:1 under the random breakage model. The mean C_s values for the three sets were 0.5929, 0.6084 and 0.5948, respectively, and the 95% confidence intervals were (0.5056,0.6742), (0.5281, 0.6854) and (0.5169, 0.6742), respectively. All datasets simulated under the random breakage model showed a significant decrease of sidedness from the ancestral level ($p<0.0001$). In contrast, the C_s value calculated for *C. reinhardtii* increased significantly to 0.8710 (Figure 3A), greatly exceeding the sidedness that would be expected in a genome that had undergone this much evolutionary change relative to its ancestor. Simulations using inferred ancestral genomes for land plant lineages (e.g. *N. tabacum*) also strongly reject the null hypothesis of random rearrangements (results not shown).

Given the large number of rearrangements observed in the *C. reinhardtii* lineage, C_f was also predicted to decrease significantly under the random breakage model, but C_f did not decrease as observed in *C. reinhardtii* (Figure 3B). The simulations with three models described above (all inversions, a small fraction of transpositions, and equal inversions and transpositions) all yielded a large decrease in clustering as expected (the observed C_f in *C. reinhardtii* is 0.03397, and the 95% confidence intervals for C_f in simulated genomes were 0.00744-0.01401, 0.00812-0.014299 and 0.0750-0.01418, respectively). When transposition was included in simulations, decreases of C_f were on a similar scale to the inversion-only simulations. Taken together, these results indicate that the remarkable increase in sidedness and functional clustering observed in *C. reinhardtii* cpDNA has not been the outcome of solely chance events. Instead, the strong deviation from the range of outcomes expected under various random breakage models implies that the genome structure is the outcome of a directional selective process.

The increased level of organization in *C. reinhardtii* cpDNA was associated with both maintenance of ancestral clusters and growth of new clusters. There were six conserved blocks containing 19 of the 85 genes shared between the *C. reinhardtii* and the *C. vulgaris* cpDNAs. These blocks include concentrations of genes from a single functional category, such as ribosomal proteins (*rpl*23-*rpl*12-*rps*19, *rpl*16-*rpl*14-*rps*8), Photosystem

II (*psb*L-*psb*F, *psb*B-*psb*T-*psb*N-*psb*H), translation apparatus (*rrn*16- *trn*I-GAU - *trn*A-UGC -*rrn*23-*rrn*5), and ATP synthase subunits (*atp*F-*atp*H). Moreover, a number of small clusters of functionally related genes inferred in the ancestral genome were brought together in *C. reinhardtii* ("rearranged clusters" in Figure 4B). These include transcription/translation genes (*trn*H-M-F; *rpl*/*rps*; *rps*3-*rpo*C2), electron transport genes (*pet*A-*pet*D), and photosynthetic genes (*psb*D-*psa*A exon 2-*psb*J) (Figure 4B). The new clusters contributed to the increase of C_f in the *C. reinhardtii* chloroplast genome.

Coordinated expression of genes in functional clusters

Co-transcription of several clusters shown in Figure 4B has been previously documented, including *psb*D-*psa*A exon 2-*psb*J-*atp*I [28], *psb*F-*psb*L [29], *pet*A-*pet*D [30], and *psb*M-*psb*Z [31]. Co-transcription of *rpl* and *rps* genes has been found in land plant chloroplasts [32]. We documented co-transcription for an additional novel functional cluster, shown in Figure 4A. Using RNA gel blots, tricistronic transcripts of *rpl*36-*rpl*23-*rpl*2 and possibly dicistronic *rpl*2-*rps*19 species could be detected. Taken together, it appears that the clusters of functionally related genes observed in *C. reinhardtii* cpDNA may be frequently co-transcribed.

Discussion

By reconstructing the possible ancestral gene order in chloroplast genomes and simulating rearrangements, we have been able to formally test and reject the null hypothesis that *C. reinhardtii* cpDNA has evolved through random rearrangements. Instead, we found that its observed gene order deviates strongly from the degree of sidedness and clustering expected under a random breakage model. *Euglena gracilis* cpDNA also has a high degree of sidedness [33], however, the asymmetry of its coding strand is concentrated in one half of the genome and associated with GC content, which could be influenced by asymmetrical replication of the chromosome [33]. In *C. reinhardtii*, the sidedness is not associated with GC content and we hypothesize that it is driven by co-transcription of genes in a functional cluster. Whereas some clusters of co-transcribed genes (e.g. *rp*l23-*rpl*2-*rps*19, *rpl*16-*rpl*14-*rps*8) were maintained in both *C. reinhardtii* and *C. vulgaris*, novel clusters clearly formed in the *C. reinhardtii* lineage (Figure 4B).

Co-transcription of neighboring genes in the *C. reinhardtii* chloroplast is a widely documented phenomenon. We demonstrated that in addition to the ribosomal protein clusters, global analyses support the elevated level of clustering of other functionally related genes. The aggregate of genes in clusters include most essential genes involved in translation and transcription, and some photosynthetic genes. Coordinated transcription may play a crucial role in the regulation of plastid gene expression in response to light or circadian rhythms [34, 35]. It is also possible that some clusters contain *cis*-elements, similar to the artificial polydeoxyadenosine sequences [36], which enhance transcription efficiency. Moreover, most of the putative co-transcription units are not conserved across chlorophytes. Therefore, the majority of functional clusters observed in *C. reinhardtii* represent new gene arrangements.

In the chloroplast gene order phylogeny (Figure 2B), the *C. reinhardtii* lineage resides on a long branch compared to the *C. vulgaris* lineage, and both genomes are more rearranged than that of *N. olivacea*, relative to the common ancestral genome of the three chlorophyte lineages. The elevated rate of chloroplast genome rearrangement in *C. reinhardtii* is associated with invasion of SDRs, which heavily populate the non-coding regions, increasing the total length of the intergenic regions compared to *C. vulgaris* cpDNA by one-third [16]. Although simple sequence repeats are common to microbial genomes [37], such elements are rare in most sequenced chloroplast genomes. Within the *Chlamydomonas* genus (Chlorophyceae), *C. reinhardtii* and *C. gelatinosa* cpDNAs exhibit a prevalence of repetitive DNA and a high degree of gene order variation compared to the *C. moewusii*/*C. pitschmannii* lineage [15, 38, 39]. The sister lineage to *C. reinhardtii* in our study, *C. vulgaris* (Trebouxiophyceae), contains numerous cpDNA repeat sequences. Besides chlorophyte algae, members of angiosperm families, including Campanulaceae [40], Fabaceae [41, 42] and Geraniaceae [43], also contain repeat elements in rearranged cpDNAs, albeit of a much lower copy number [40-43]. These repeat elements may act as molecular "grease" that facilitates non-homologous recombination and creates a pool of diverse genome structures subject to selective retention. Future investigations will test whether the increased rates of rearrangement in plastid genomes with dispersed repeats typically lead to increased sidedness and functional clustering as we infer for *C. reinhardtii.*

Gene order changes reflect relatively rare evolutionary events and are expected to result in much less homoplasy than substitution events in nucleotide or protein sequences over a deep time scale [44]. Phylogeny reconstruction using GRAPPA is highly accurate even for divergent genomes [45], and thus the ancestral gene orders inferred in our study contained sufficient phylogenetic information. The only other software for genome rearrangement phylogeny, BADGER [46], performed poorly on this data set (results not shown). GRAPPA usually inferred unique ancestral gene orders on many data sets we tested. Furthermore, analyses on simulated data have shown that the inferred gene orders scored almost as well as true ancestral gene orders [47]. In our simulation tests of three genomes with 85 genes each, and branch lengths of 50, 20 and 20 (roughly corresponding to the branches leading to *C. reinhardtii*, *C. vulgaris* and *N. olivacea*; see Methods), the average score for ancestral gene orders computed by GRAPPA was only about 7% less than the true scores. In practice, we observed that the less optimal gene orders generally required more rearrangements. Therefore, it is quite likely that any error in our estimation of ancestral gene order has resulted in a downward bias in the inferred number of rearrangements on the branch leading to *C. reinhardtii*. Increasing the number of rearrangements on this branch would only lead to a more certain rejection of the neutrality of rearrangements.

The accuracy of ancestral genome reconstruction also depends on the degree of divergence among extant taxa and taxon sampling. For example, accurate reconstruction of ancestral genomes at the mammalian CFTR locus was achieved at the DNA level [48]. The high-quality reconstruction was attributed to a dense sampling of syntenic genome sequences from eutherian mammals, and the lack of gene order rearrangement at the locus. Because the *C. reinhardtii* cpDNA is one of the most rearranged chloroplast genomes sequenced to date, we included all available chlorophyte chloroplast genomes for evolutionary distance estimation and ancestral gene order reconstruction. The accuracy of our ancestral gene order estimation may improve with inclusion of additional chlorophyte plastid gene orders as they become available, but we do not foresee a substantial reduction in the inferred number of rearrangements separating *C. reinhardtii* and *C. vulgaris* from their common ancestor.

Inversions are thought to be much more common than transpositions in chloroplast genome evolution [27], and our estimation of ancestral genome order was made with the assumption that all rearrangements were inversions. However, we did consider the contribution of inversions and transpositions under different scenarios in the simulation from the ancestral genome. It should be noted that there is not a unique phylogeny distance measure using transposition only, because computationally one transposition is equivalent to two or three inversions [49]. For this reason, we designed our simulations to allow for various ratios of inversion and transposition events. The results of our simulation study did not vary significantly under these scenarios.

The GRAPPA-IR algorithm was developed to account for the inverted repeat (IR) region found in most plastid genomes The IR region seems to evolve at a slower rate in both nucleotide sequence and gene order than the single copy regions [50], and frequent intra-molecular recombination homogenizes the two copies [6, 51]. The most conserved gene set in the IR region is the rRNA operon. In IR-containing green plastids, the order of rRNA genes is conserved, but the IR boundaries can vary greatly even within one genus [52]. The IR may restrict rearrangements that cross the boundary of single copy regions, and thus concentrate gene order changes within single copy regions. However, this hypothetical constraint of the IR on genome rearrangements seems to have been lost in the *C. reinhardtii*/*C. vulgaris* lineage. Notably, both lineages have undergone extensive rearrangements since their divergence from a common ancestor, and they contain only a few conserved clusters encoding rRNA or ribosomal proteins. In either genome, genes that typically reside together in the LSC region have often been scrambled and scattered. When comparing the ancestral genome to the *C. vulgaris* gene order, there was no distinction of LSC and SSC regions although many large clusters were still shared (additional file 4). If there were constraints on the breakpoint locations, as experimentally identified in bacterial inversion mutants [53], it would limit the possible paths of evolution, and these constraints on the ancestral gene orders would increase the number of rearrangements relative to the estimations derived from GRAPPA. Therefore, as discussed above, our approach of detecting strong deviation from expectation is conservative in that the number of rearrangements may be underestimated.

Recent studies of plant, animal and fungal genomes have shown that genes involved in the same pathways or genes sharing similar expression patterns are often spatially clustered [1, 5, 54]. In eukaryotes, the operon structure has only been demonstrated in the nematode *Caenorhabditis* [55]. Comparative analyses of yeast genomes indicate that rearrangements brought together duplicate genes forming the *DAL* cluster involved in allantoin metabolism [56]. In this study, we demonstrated that positive selection for increased clustering has influenced gene order in the chloroplast. Gene clusters, as opposed to separated genes, permit polycistronic transcription and thus fewer transcriptional regulation units. Co-transcription may be facilitated by close spacing of genes in cpDNA because transcription termination is inefficient [57]. Although post-transcriptional RNA processing often creates multiple single-gene transcripts, co-transcription foments an initial stoichiometric accumulation of RNA corresponding to each gene in a cluster. Thus, large clusters can be advantageous in coordinating gene expression on this level.

Experimental approaches are necessary to understand whether these gene clusters function as operons. Because chloroplast primary transcripts are heavily processed – as just one example, the *psb*B cluster in maize accumulates as at least 15 distinct mRNA species with varying translational capacities [58] – direct analysis of the functional advantages of clustering in chloroplasts is challenging. Indeed, *Chlamydomonas* may be a special case, since unlike land plants it has a single rather than multiple RNA polymerases [35]. This situation does not allow differential expression by promoter selectivity, and may therefore serve as a selective force that favors physical grouping of genes rather than evolution of promoter sequences of dispersed genes.

Conclusions

In conclusion, we infer that gene order in the *C. reinhardtii* plastid evolved in a non-random fashion, and hypothesize that genome structure has been influenced by directional selection acting on variation generated by an increased rate of rearrangement. Our results provide strong evidence that genetic responses to natural selection occur at the level of genome organization. By estimating the ancestral gene order and simulating rearrangements under a null model, we provide a formal demonstration that the chloroplast genome of *C. reinhardtii* has been shaped by natural selection. Although the model of natural selection on gene order remains to be developed, application of our methods to sequences of additional chlorophyte plastid genomes would help to improve the accuracy of the ancestral genome reconstruction and inferred branch lengths. The complex process of gene duplication and loss in bacterial and eukaryotic nuclear genomes presents challenges to reconstruction of ancestral gene orders. Still, the development of new comparative tools [59] gives us hope that the type of analysis presented in this paper will soon be applicable to eukaryotic genomes.

Methods

Functional clustering of chloroplast genes

We defined a “functional cluster” as contiguous genes encoded on one strand from one of the following categories: transcription/translation, photosystem I and II, electron transport (cytochrome b6/f complex), and ATP synthase (See additional file 1).

Kolmogorov-Smirnov test of random clusters

A random cluster consists of genes from any functional category. The n=85 genes shared in the seven chloroplast genomes shown in Figure 1 were divided into 11 equal sized blocks of r_j=7 genes and one block of 8 genes so that the block sizes and number of blocks are equal. If m_{ij} genes were from the functional category i (total T_i genes) in the jth block, the observed cumulative frequency was $u_i = \sum_i m_{ij} \big/ r_j$. The Kolmogorov-Smirnov test measures the deviation of the observed u_i from the expected from the random breakage model [13]. The test statistic D_n was calculated for each functional category separately.

$$D_n = \max[\max(\frac{T_i}{n} - u_i), \max(u_i - \frac{T_i - 1}{n})]$$

Phylogeny of chloroplast genomes

Alignments of 50 proteins shared in the 8 chloroplast genomes shown in Figure 2A were concatenated into one data matrix (Additional file 2). 1,000 bootstrap replicates were conducted on the data set using PAUP* 4.0b10 with maximum parsimony and using MEGA with neighbor-joining methods and the Poisson-corrected distance. Maximum likelihood analysis with 100 bootstrap replicates was performed using PHYLIP3.6 with

JTT distance and gamma = 0.5. GRAPPA was not used to construct the reference phylogeny.

Inferring ancestral gene orders

The ancestral gene order was inferred from the gene orders of extant genomes on the best-scored tree following two steps. First, the gene contents for the LSC, SSC and IR regions of ancestral genomes of IR-containing cpDNAs were inferred based on parsimony. Changes in gene copy number due to IR expansion or contraction were considered the last step of gene order changes, and thus the gene contents of ancestral genomes were determined. The ancestral gene orders on the phylogeny for five genomes (excluding *C. vulgaris* and *C. reinhardtii*) were computed using GRAPPA-IR [25], which is a modified version of GRAPPA that scores rearrangements independently within LSC, SSC or IR. Second, the chlorophyte algal gene orders (the extant chloroplast gene orders of *N. olivacea*, *C. reinhardtii* and *C. vulgaris* and the inferred ancestral genome of *N. olivacea* from step one) and the gene order of *M. viride* were used for the inference of the common ancestral gene order of *C. vulgaris* and *C. reinhardtii*. The data set contains duplicated *trn*V-UAC and *trn*G-GCC in *C. vulgaris*, *trn*E-UUC and *psb*A in *C. reinhardtii* and three trans-splicing *psa*A exons in C. *reinhardtii*. The IR regions contained rRNA genes in the same order and orientation in each genome except that one copy was lost in the lineage leading to *C. vulgaris*. To score the genomes with gene duplications and deletions, multiple data sets were created each containing genomes with equalized gene contents by the following assignment rules: one copy of each duplicate genes outside the typical IR was chosen; the IR region lost in *C. vulgaris* was inserted to all possible locations in that genome. Preferably, we should test all these datasets (3,936 total) with inversion medians; however, such computation on one dataset alone will take more than a month. To overcome this limitation, these datasets were computed using breakpoint medians, and the assignment yielded the shortest tree was chosen for a full evaluation by GRAPPA. Because the gene contents of LSC and SSC in *C. reinhardtii* were different from other chloroplast genomes in the study, we allows free rearrangements such that genes in LSC or SSC could commute across the IR.

Ancestral gene order simulation

A set of simulation experiments were conducted to evaluate the accuracy of ancestral genome reconstruction with long branches. Three genomes with 85 genes each were generated from a defined ancestral gene order, and the branch lengths (inversion distances) were 50, 20 and 20, respectively. The true gene order score was 90 (equals the tree length). The scores were computed for inferred ancestral gene orders by GRAPPA using inversion medians and the random breakage model and then compared to the true score. The experiment was repeated on 30 data sets.

Random genome rearrangement simulation

Gene orders were simulated under the assumption that the rearrangements involve random breakpoints placed between genes. Initial gene orders were set based on the inferred ancestral gene orders estimated. Random rearrangement operations on the initial genomes were performed for the number of replicates according to the number of rearrangements inferred by GRAPPA. The parameters input to the model were the ratios of inversion and transposition (1:0, 10:1, 1:1) to test the sensitivity of the findings to the specific rearrangement model. The simulated genomes had identical gene content but scrambled gene orders relative to those observed in extant genomes, with the exception that inverted repeats were maintained. Test statistics (below) were calculated for each simulated replicate of 10,000 total and the frequency distributions were used to test the null hypothesis of random rearrangement.

Sidedness index (C_s)

We designed the sidedness index (C_s) to measure the degree to which neighboring genes are clustered on the same strand (side) of the chromosome. A "sided block" includes only adjacent genes on one strand, and the number of sided blocks in a genome is designated as n_{SB}, while the total number of genes is n. C_s is defined as

$$C_s = (n-n_{SB})/(n-1).$$

When C_s reaches the maximum of 1, all genes are located on one side. If every gene resides on the strand opposite its neighbors, C_s approaches a minimum of zero.

Functional cluster index (C_f)

We divided a genome of total n genes to J sided blocks ($r_1, r_2, \ldots r_J$). In a block, we assigned genes to functional categories. Let the numbers of genes in the ith functional category and the jth block be m_{ij}, the functional cluster index C_f is

$$C_f = \frac{1}{J}\sum_{j=1}^{J}\frac{r_j}{n}\sum_{i=1}^{4}\binom{m_{ij}}{2}\Big/\binom{r_j}{2}.$$

A larger value of C_f indicates that functionally related genes are more clustered into blocks.

RNA analysis

Wild-type CC-124 cells were grown in Tris-Acetate-Phosphate medium [60] under continuous light to mid-log phase. RNA was isolated from 10 mL of cells as previously described [61]. For filter hybridization, 5 μg of total RNA was fractionated in 1.2%

agarose and 6% formaldehyde gels, transferred to nylon membranes, and probed with gene-specific PCR products labeled by random priming according to Church and Gilbert [62].

List of Abbreviations

cpDNA, chloroplast DNA; IR, inverted repeat; SDR, short dispersed repeat

Authors' contributions

LC conducted the analysis and drafted the manuscript. JLM and CWD conceived the study, helped with the analyses and contributed to the text. LSW contributed the code for the genome simulator. JT carried out the ancestral genome reconstruction. LR conducted the RNA analysis. DBS provided further experimental data review and revision of the draft. All authors read and approved the final manuscript.

Acknowledgements

We thank A. Jarosz, H. Ma, J. Marden, W. Martin, W. Miller, and D. Schemske for valuable suggestions and comments. This work was supported by NSF awards DBI 0115684 and DEB 0120709 to CWD. JT is supported by the University of South Carolina and part of the work was done while he was visiting the National Evolutionary Synthesis Center. *Chlamydomonas* genomics work at BTI was supported by NSF awards MCB 9975765 and MCB 0091020 to DBS.

References

1. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order**. *Nat Rev Genet* 2004, **5**(4):299-310.
2. Kallas T, Spiller S, Malkin R: **Primary structure of cotranscribed genes encoding the Rieske Fe-S and cytochrome f proteins of the cyanobacterium *Nostoc* PCC 7906**. *Proc Natl Acad Sci U S A* 1988, **85**(16):5794-5798.
3. Stoebe B, Kowallik KV: **Gene-cluster analysis in chloroplast genomics**. *Trends Genet* 1999, **15**(9):344-347.
4. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome**. *Nat Genet* 2002, **31**(2):180-183.
5. Lee JM, Sonnhammer EL: **Genomic gene clustering analysis of pathways in eukaryotes**. *Genome Res* 2003, **13**(5):875-882.

6. Palmer JD: **Evolution of chloroplast and mitochondrial DNA in plants and algae**. In: *Molecular Evolutionary Genetics.* Edited by MacIntyre RJ. New York: Plenum Press; 1985: 131-240.
7. Pancic PG, Strotmann H, Kowallik KV: **Chloroplast ATPase genes in the diatom *Odontella sinensis* reflect cyanobacterial characters in structure and arrangement**. *J Mol Biol* 1992, **224**(2):529-536.
8. Nadeau J, Taylor BA: **Length of chromosome segments conserved since divergence of man and mouse**. *Proc Natl Acad Sci USA* 1984, **81**:814-818.
9. Cosner ME, Jansen RK, Moret BM, Raubeson LA, Wang LS, Warnow T, Wyman S: **A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:104-115.
10. Moret BM, Wang LS, Warnow T, Wayman SK: **New approaches for reconstructing phylogenies from gene order data**. *Bioinformatics* 2001, **17**:S165-S173.
11. Tang JJ, Moret BME: **Phylogenetic reconstruction from gene-rearrangement data with unequal gene content**. In: *Lecture Notes in Computer Science.* vol. 2748. Berlin: Springer-Verlag; 2003: 37-46.
12. Tang J, Moret BM: **Scaling up accurate phylogenetic reconstruction from gene-order data**. *Bioinformatics* 2003, **19 Suppl 1**:i305-312.
13. Pevzner P, Tesler G: **Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution**. *Proc Natl Acad Sci USA* 2003, **100**:7672-7677.
14. Trinh P, McLysaght A, Sankoff D: **Genomic features in the breakpoint regions between syntenic blocks**. *Bioinformatics* 2004, **20 Suppl 1**:I318-I325.
15. Boudreau E, Turmel M: **Extensive gene rearrangements in the chloroplast DNAs of *Chlamydomonas* species featuring multiple dispersed repeats**. *Mol Biol Evol* 1996, **13**(1):233-243.
16. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB: ***Chlamydomonas* chloroplast chromosome: islands of genes in a sea of repeats**. *Plant Cell* 2002, **14**:2659-2679.
17. Wakasugi T, Nagai T, Kapoor M, Sugita M, Ito M, Ito S, Tsudzuki J, Nakashima K, Tsudzuki T, Suzuki Y, Hamada A, Ohta T, Inamura A, Yoshinaga K, Sugiura M: **Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division**. *Proc Natl Acad Sci U S A* 1997, **94**(11):5967-5972.
18. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchishinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M: **The Complete Nucleotide-Sequence of the Tobacco Chloroplast Genome - Its Gene Organization and Expression**. *Embo Journal* 1986, **5**(9):2043-2049.
19. Ohyama K, Fukuzawa H, Kohchi T, Sano T, Sano S, Shirai H, Umesono K, Shiki Y, Takeuchi M, Chang Z, et al.: **Structure and organization of *Marchantia polymorpha* chloroplast genome. I. Cloning and gene identification**. *J Mol Biol* 1988, **203**(2):281-298.

20. Turmel M, Otis C, Lemieux C: **The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants**. *Proc Natl Acad Sci U S A* 2002, **99**(17):11275-11280.

21. Turmel M, Otis C, Lemieux C: **The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes**. *Proc Natl Acad Sci U S A* 1999, **96**(18):10248-10253.

22. Lemieux C, Otis C, Turmel M: **Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution**. *Nature* 2000, **403**(6770):649-652.

23. Loffelhardt W, Bohnert HJ, Bryant DA: **The cyanelles of *Cyanophora paradoxa***. *Crit Rev Plant Sci* 1997, **16**(4):393-413.

24. Karol KG, McCourt RM, Cimino MT, Delwiche CF: **The closest living relatives of land plants**. *Science* 2001, **294**(5550):2351-2353.

25. Cui L, Tang J, Moret BME, dePamphilis CW: **Inferring ancestral chloroplast genomes with duplications**. In: *TR-CS-2005-08.* University of New Mexico; 2005.

26. Palmer JD: **Contrasting modes and tempos of genome evolution in land plant organelles**. *Trends Genet* 1990, **6**(4):115-120.

27. Boudreau E, Turmel M: **Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat**. *Plant Mol Biol* 1995, **27**(2):351-364.

28. Choquet Y, Goldschmidt-Clermont M, Girard-Bascou J, Kuck U, Bennoun P, Rochaix JD: **Mutant phenotypes support a trans-splicing mechanism for the expression of the tripartite psaA gene in the *C. reinhardtii* chloroplast**. *Cell* 1988, **52**:903-914.

29. Mor TS, Ohad I, Hirschberg J, Pakrasi HB: **An unusual organization of the genes encoding cytochrome b559 in *Chlamydomonas reinhardtii*: psbE and psbF genes are separately transcribed from different regions of the plastid chromosome**. *Mol Gen Genet* 1995, **246**(5):600-604.

30. Sturm NR, Kuras R, Buschlen S, Sakamoto W, Kindle KL, Stern DB, Wollman FA: **The petD gene is transcribed by functionally redundant promoters in *Chlamydomonas reinhardtii* chloroplasts**. *Mol Cell Biol* 1994, **14**:6171-6179.

31. Higgs DC, Kuras R, Kindle KL, Wollman FA, Stern DB: **Inversions in the *Chlamydomonas* chloroplast genome suppress a petD 5' untranslated region deletion by creating functional chimeric mRNAs**. *Plant J* 1998, **14**(6):663-671.

32. Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K: **RNA editing in hornwort chloroplasts makes more than half the genes functional**. *Nucleic Acids Res* 2003, **31**(9):2417-2423.

33. Morton BR: **Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis***. *Proc Natl Acad Sci USA* 1999, **96**:5123-5128.

34. Thompson RJ, Mosig G: **Light affects the structure of *Chlamydomonas* chloroplast chromosomes**. *Nucleic Acids Res* 1990, **18**:2625-2631.

35. Eberhard S, Drapier D, Wollman FA: **Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number,**

transcription, transcript abundance and translation rate in the chloroplast of *Chlamydomonas reinhardtii*. *Plant J* 2002, **31**(2):149-160.

36. Lisitsky I, Rott R, Schuster G: **Insertion of polydeoxyadenosine-rich sequences into an intergenic region increases transcription in *Chlamydomonas reinhardtii* chloroplasts**. *Planta* 2001, **212**(5-6):851-857.

37. Saunders NJ, Peden JF, Hood DW, Moxon ER: **Simple sequence repeats in the *Helicobacter pylori* genome**. *Mol Microbiol* 1998, **27**(6):1091-1098.

38. Lemieux B, Turmel M, Lemieux C: **Chloroplast DNA variation in *Chlamydomonas* and its potential application to the systematics of this genus**. *Biosystems* 1985, **18**(3-4):293-298.

39. Boudreau E, Otis C, Turmel M: **Conserved gene clusters in the highly rearranged chloroplast genomes of *Chlamydomonas moewusii* and *Chlamydomonas reinhardtii***. *Plant Mol Biol* 1994, **24**(4):585-602.

40. Cosner ME, Jansen RK, Palmer JD, Downie SR: **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* 1997, **31**(5):419-429.

41. Perry AS, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH: **Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement**. *DNA Res* 2002, **9**(5):157-162.

42. Milligan BG, Hampton JN, Palmer JD: **Dispersed repeats and structural reorganization in subclover chloroplast DNA**. *Mol Biol Evol* 1989, **6**(4):355-368.

43. Price RA, Calie PJ, Downie SR, Logsdon J, J.M., Palmer JD: **Chloroplast DNA variation in the Geraniaceae - a preliminary report**. In: *Proc Int Geraniaceae Symp.* Edited by Vorster P. Monvilla, South Africa: University of Stellenbosch; 1990: 235-244.

44. Rokas A, Holland PW: **Rare genomic changes as a tool for phylogenetics**. *Trends Ecol Evol* 2000, **15**(11):454-459.

45. Wang LS, Jansen, R. K., Moret, B. M., Raubeson, L. A., Warnow, T.: **Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study**. In: *Pac Symp Biocomput.* River Edge: World Scientific Pub.; 2002: 524-535.

46. Simon D, Larget B: **Bayesian Analysis to Describe Genomic Evolution by Rearrangement (BADGER), version 1.01 beta**. 2004.

47. Siepel AC, Moret BME: **Finding an Optimal Inversion Median: Experimental Results**. In: *Lecture Notes in Computer Science.* vol. 2149: Springer-Verlag; 2001: 189-203.

48. Blanchette M, Green ED, Miller W, Haussler D: **Reconstructing large regions of an ancestral mammalian genome in silico**. *Genome Res* 2004, **14**(12):2412-2423.

49. Wang L-S: **Exact-IEBP: A New Technique for Estimating Evolutionary Distances between Whole Genomes**. In: *Lecture Notes in Computer Science.* vol. 2149: Springer-Verlag; 2001: 175-188.

50. Wolfe KH, Li WH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs**. *Proc Natl Acad Sci U S A* 1987, **84**(24):9054-9058.
51. Lemieux B, Turmel M, Lemieux C: **Recombination of *Chlamydomonas* Chloroplast DNA Occurs More Frequently in the Large Inverted Repeat Sequence Than in the Single-Copy Regions**. *Theor Appl Genet* 1990, **79**(1):17-27.
52. Goulding SE, Olmstead RG, Morden CW, Wolfe KH: **Ebb and flow of the chloroplast inverted repeat**. *Mol Gen Genet* 1996, **252**(1-2):195-206.
53. Segall AM, Roth JR: **Recombination between homologies in direct and inverse orientation in the chromosome of *Salmonella*: intervals which are nonpermissive for inversion formation**. *Genetics* 1989, **122**(4):737-747.
54. Williams EJ, Bowles DJ: **Coexpression of neighboring genes in the genome of *Arabidopsis thaliana***. *Genome Res* 2004, **14**(6):1060-1067.
55. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, Kim SK: **A global analysis of *Caenorhabditis elegans* operons**. *Nature* 2002, **417**(6891):851-854.
56. Wong S, Wolfe KH: **Birth of a metabolic gene cluster in yeast by adaptive gene relocation**. *Nat Genet* 2005, **37**(7):777-782.
57. Monde RA, Schuster G, Stern DB: **Processing and degradation of chloroplast mRNA.** *Biochimie* 2000, **82**:573-582.
58. Barkan A: **Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic RNAs.** *EMBO J* 1988(7):2637-2644.
59. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes**. *Proc Natl Acad Sci U S A* 2003, **100**(20):11484-11489.
60. Harris EH: **The *Chlamydomonas* sourcebook: A comprehensivve guide to biology and laboratory use**. San Diego: Academic Press; 1989.
61. Drager RG, Higgs DC, Kindle KL, Stern DB: **5' to 3' exoribonucleolytic activity is a normal component of chloroplast mRNA decay pathways**. *Plant J* 1999, **19**(5):521-531.
62. Church GM, Gilbert W: **Genomic sequencing**. *Proc Natl Acad Sci U S A* 1984, **81**(7):1991-1995.

Figure legends

Figure 1 - Extensive rearrangement in *Chlamydomonas reinhardtii* and *Chlorella vulgaris* cpDNAs.

Representative cpDNAs from land plants and green algae are arranged to reflect their phylogenetic relationships. The scale bar indicates 20 kb. Each genome is linearized and drawn as a grey bar. Genes are drawn as colored rectangles and with those encoded on the positive strand above the genome bar. Colored lines connect the homologs included in this study and the functional category is shown by specific colors.

Figure 2 - The phylogeny of cpDNAs.

(A) The cpDNA phylogeny based on analysis of 50 concatenated proteins. The phylogeny includes major green plant and algal lineages and the outgroup *Cyanophora*. The bootstrap support values from maximum parsimony/neighbor joining/maximum likelihood analyses are labeled near each node. (B) Estimated inversion distances considering 85 common genes on the cpDNA phylogeny. There is an increase of rearrangements on branches leading to *C. reinhardtii*. and *C. vulgaris*, from a common ancestor indicated by an arrow.

Figure 3 - Comparison of sidedness and functional cluster indices in *C. reinhardtii* cpDNA to those of simulated genomes

(A)The sidedness index C_s observed in *C. reinhardtii* (indicated by an arrow) is significantly larger than C_s of gene orders simulated under the random breakage model (inversion only) and the estimated ancestral genome indicated in Figure 2B. (B) The functional cluster index C_f for *C. reinhardtii* (indicated by a solid horizontal line) is greater than that for the inferred ancestral genome (dashed line), in contrast to the decrease predicted by three sets of simulations under the random breakage model. Models 1, 2 and 3 specified the inversion/transposition ratios to be 1:0, 10:1 and 1:1, respectively, in simulations with 10,000 replicates. The box section of the box plot indicates the first quartile, median and the third quartile of the distribution.

Figure 4 - Selected functional clusters from *C. reinhardtii* cpDNA.

(A)Evidence for co-transcription of the genes *rpl*36-*rpl*23-*rpl*2-*rps*19. The gel was loaded with total RNA from wild-type cells, and shows new evidence for co-transcription (see text). The top left lane is an over-exposure of the *rpl*36 gel. Transcripts 1 and 2 (3.5 and

3.3 kb) are tricistronic *rpl*36-*rpl*23-*rpl*2, transcript 3A (2.5 kb) is *rpl*36-*rpl*23, and transcript 3B (2.5 kb) is probably *rpl*2-*rps*19. Single gene transcripts are labeled "mono". (B)Rearranged functional clusters, which were absent from the inferred common ancestor of *C. reinhardtii* and *C. vulgaris*, were identified in *C. reinhardtii* (genes connected by bold black lines). Cyan lines connect conserved clusters retained from the ancestor cpDNA. The genes are displayed in the coding direction, and from top to bottom relative to their order in the genome. The exception is *psb*N, which is on the opposite strand relative to other genes shown (*psb*T-B-N-H). A scale bar of 1 kb is shown below and at the left of each gene cluster.

Tables

Table 1 - The Kolmogorov-Smirnov test of gene clustering by the functional category in cpDNAs §

cpDNA	D_n (p-value)			
	Translation and transcription	Photosystem I and II	Electron Transport	ATP synthase
Chlorella	0.214(.6418)	0.488(.0066)	0.750(.0000)	0.833(.0000)
Chlamydomonas	0.198(.6866)	0.473(.0060)	0.780(.0000)	0.769(.0000)
Nephroselmis	0.209(.6207)	0.484(.0046)	0.703(.0000)	0.846(.0000)
Mesostigma	0.275(.2786)	0.549(.0008)	0.769(.0000)	0.846(.0000)
Chaetosphaeridium	0.242(.4388)	0.484(.0046)	0.714(.0000)	0.846(.0000)
Marchantia	0.341(.0986)	0.473(.0060)	0.714(.0000)	0.846(.0000)
Nicotiana	0.264(.3283)	0.473(.0060)	0.769(.0000)	0.846(.0000)

§The test statistic D_n measures whether the distribution of functionally related genes is random in gene clusters. Total 85 shared genes between seven cpDNAs were included.

Additional files

Additional file 1 – Gene coding and functional categories.
Text file, lists the names of 85 genes included in the study and corresponding functional categories.

Additional file 2 – Protein alignment matrix.
Text file, with a NEXUS format data matrix of concatenated proteins from seven chloroplast genomes and the outgroup, *Cyanophora paradoxa*.

Additional file 3 – The gene order data set.
Text file, contains gene orders of seven chloroplast genomes, computed C_s and C_f indices, and the inferred rearrangement phylogeny.

Additional file 4 – Comparison of gene clusters.
Text file, shows gene clusters shared between the inferred ancestral genome of *C. reinhardtii* and *C. vulgaris* to the cpDNA of *C. vulgaris* and *N. olivacea*.

Additional file 5 – Inversions separating *N. tabacum* and *M. polymorpha* cpDNA.
Text file, shows the possible scenarios to transform the chloroplast gene order of *N. tabacum* to *M. polymorpha* cpDNA through inversions.

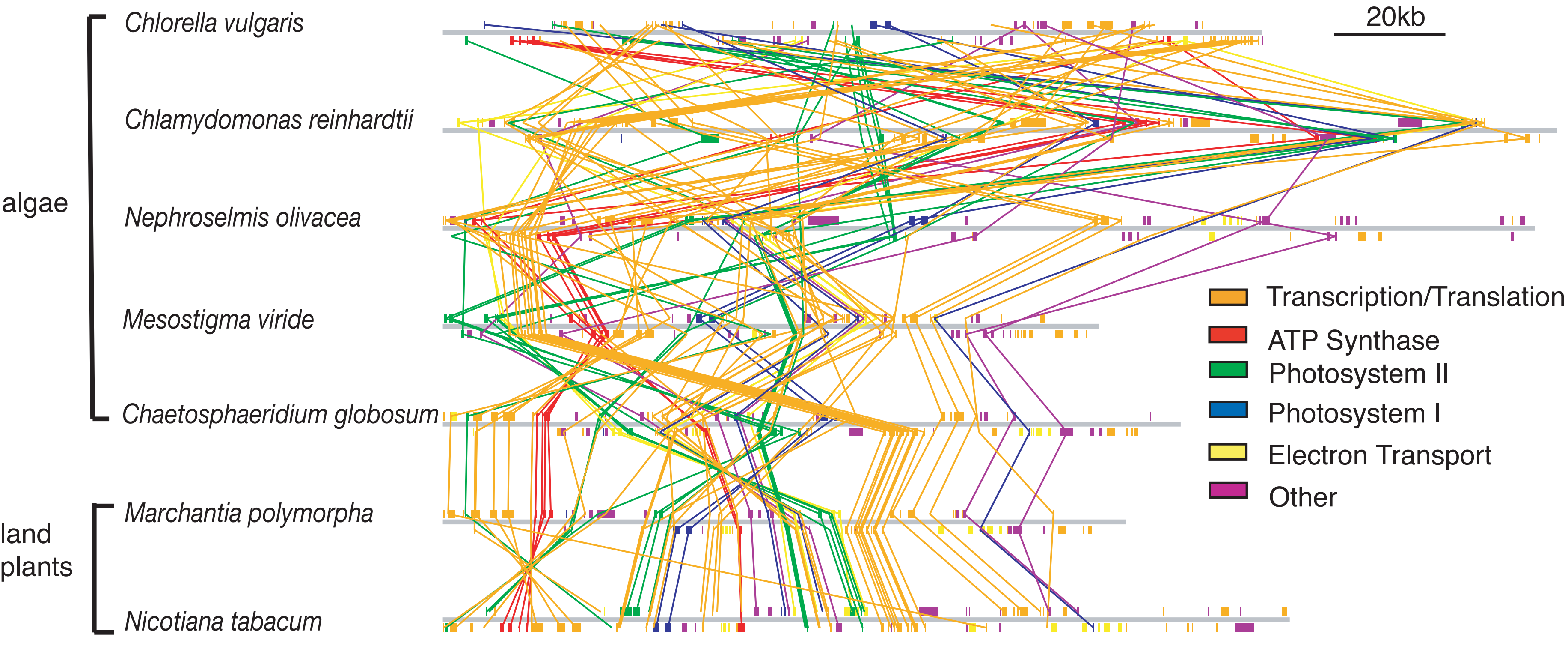

Figure 1

Figure 1

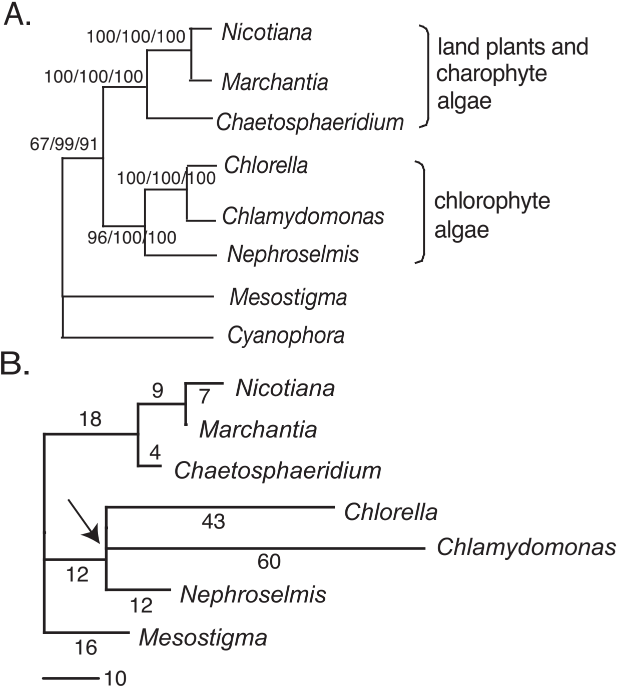

Figure 2

Figure 2

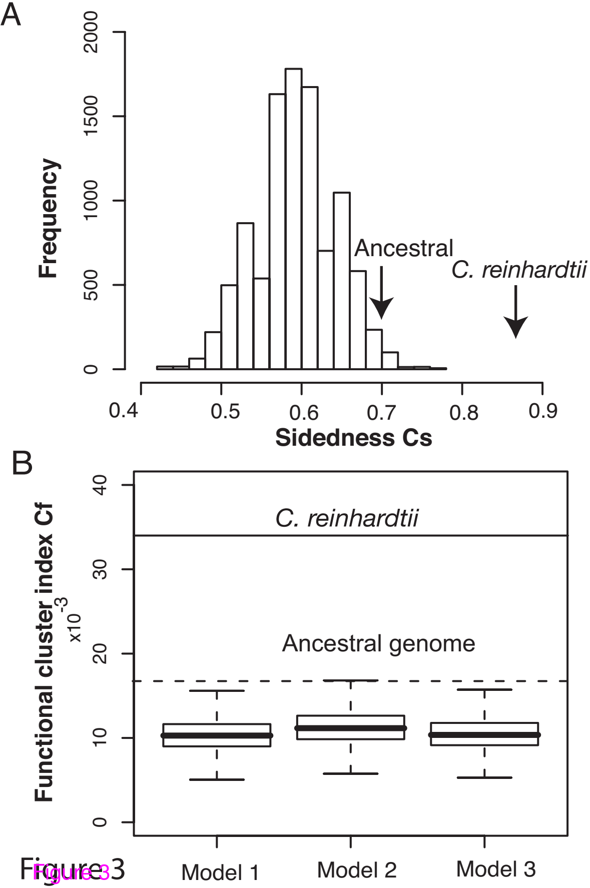

Figure 3

Figure 3

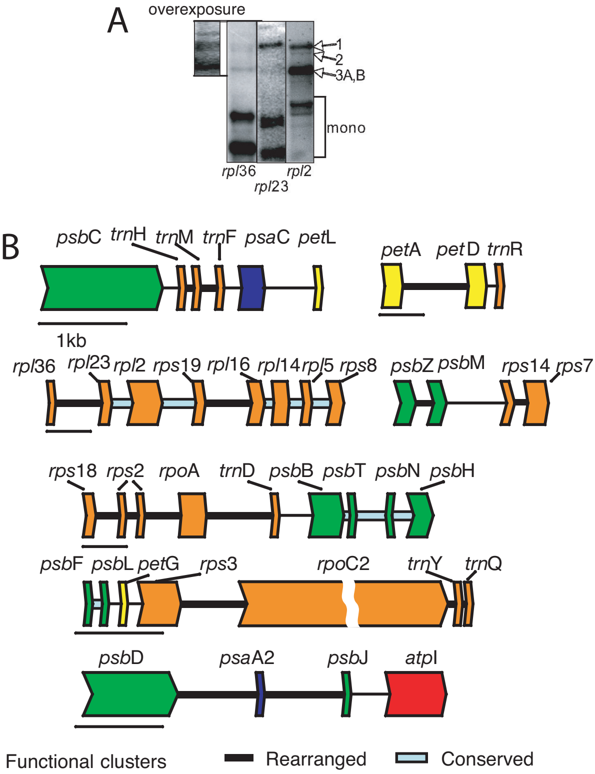

Figure 4

Figure 4

Additional files provided with this submission:

Additional file 5 : cui_addfile5.txt : 1Kb
http://www.biomedcentral.com/imedia/1002328855785186/sup5.TXT

Additional file 4 : cui_addfile4.txt : 1Kb
http://www.biomedcentral.com/imedia/1008167807851857/sup4.TXT

Additional file 3 : cui_addfile3.txt : 6Kb
http://www.biomedcentral.com/imedia/1701569914785185/sup3.TXT

Additional file 2 : cui_addfile2.txt : 169Kb
http://www.biomedcentral.com/imedia/3447299877851764/sup2.TXT

Additional file 1 : cui_addfile1.txt : 3Kb
http://www.biomedcentral.com/imedia/1931512132785176/sup1.TXT