plants

MDPI

Article

Evolutionary Trends in the Mitochondrial Genome of Archaeplastida: How Does the GC Bias Affect the Transition from Water to Land?

Joan Pedrola-Monfort [1], David Lázaro-Gimeno [1], Carlos G. Boluda [1,2], Laia Pedrola [1], Alfonso Garmendia [3], Carla Soler [4] and Jose M. Soriano [4,*]

[1] Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, 46980 Paterna, Spain; joan.pedrola@uv.es (J.P.-M.); david.lazaro-gimeno@uv.es (D.L.-G.); Carlos.Boluda@ville-ge.ch (C.G.B.); laiapedrolavidal@gmail.com (L.P.)

[2] Unité de Phylogénie et Génetique Moléculaires, Conservatoire et Jardin Botaniques, Chambésy, 1292 Geneva, Switzerland

[3] Mediterranean Agroforestry Institute, Department of Agroforest Ecosystems, Polytechnic University of Valencia, 46022 Valencia, Spain; algarsal@upvnet.upv.es

[4] Biomaterials, Institute of Materials Science, University of Valencia, 46980 Paterna, Spain; carla.soler@uv.es

* Correspondence: jose.soriano@uv.es; Tel.: +34-963-543-056

Received: 20 February 2020; Accepted: 11 March 2020; Published: 12 March 2020

Abstract: Among the most intriguing mysteries in the evolutionary biology of photosynthetic organisms are the genesis and consequences of the dramatic increase in the mitochondrial and nuclear genome sizes, together with the concomitant evolution of the three genetic compartments, particularly during the transition from water to land. To clarify the evolutionary trends in the mitochondrial genome of Archaeplastida, we analyzed the sequences from 37 complete genomes. Therefore, we utilized mitochondrial, plastidial and nuclear ribosomal DNA molecular markers on 100 species of Streptophyta for each subunit. Hierarchical models of sequence evolution were fitted to test the heterogeneity in the base composition. The best resulting phylogenies were used for reconstructing the ancestral Guanine-Cytosine (GC) content and equilibrium GC frequency (GC*) using non-homogeneous and non-stationary models fitted with a maximum likelihood approach. The mitochondrial genome length was strongly related to repetitive sequences across Archaeplastida evolution; however, the length seemed not to be linked to the other studied variables, as different lineages showed diverse evolutionary patterns. In contrast, Streptophyta exhibited a powerful positive relationship between the GC content, non-coding DNA, and repetitive sequences, while the evolution of Chlorophyta reflected a strong positive linear relationship between the genome length and the number of genes.

Keywords: Archaeplastida; GC bias; equilibrium GC frequency; GC content concomitance; mitochondrial genomic pattern

1. Introduction

Mitochondrial, plastidial and nuclear genome lengths (GL) increased dramatically due to the addition of non-coding DNA (%NC) during the evolution of green plants, particularly during the transition from water to terrestrial life. This phenomenon occurred in parallel with an increase in Guanine-Cytosine (GC) content (%GC) and organism complexity [1]. The interactions among GL, %NC, the number of repeated sequences (NRS), their total length (RSL), and the %GC are not yet well understood for any of the three genetic compartments; however, plastids appear to be the most evolutionarily stable, with few changes in GL or %GC [2]. Evolutionary changes in the aforementioned variables may have occurred concurrently at two or three genetic compartments, and

the factors determining this concurrence or the lack thereof are among the most intriguing puzzles in the evolutionary biology of primary photosynthetic eukaryotes.

Archaeplastida include Glaucophyta, Rhodophyta (red algae), and Viridiplantae (green plants), although the monophyly of this group is not exempt from controversy [3,4]. Green plants are further divided into two main clades: Chlorophyta, including most unicellular and marine algae; Streptophyta, including most freshwater algae and land plants [5]. All these lineages have three genetic compartments with well-coordinated working biochemical machinery, and they differ significantly in their architecture and evolution [6]. We can see that lateral transfer among the three compartments occurred, especially from the organelles to the nucleus and from the plastid to the mitochondrion, whereas transfer from the mitochondrion to the plastid was almost non-existent [7]. Meaningful differences can be reflected in some of the mitochondrial genome (mtDNA) characteristics in the Archaeplastida lineages, such as the GLs, genetic code, codon usage, gene content, and the degree of ribosomal gene fragmentation [8]. Throughout the transition from water to land, terrestrial plants acquired some peculiar features in their mtDNA, including large genomes with a high %NC, editing at the transcriptional level, genomic recombination, trans-splicing introns, foreign DNA insertions, lateral gene transfer, and gene duplications. These features are not yet widely studied in streptophyte green algae and early land plants [9]. Selection was tested in both Viridiplantae lineages as the driving force that increases mtDNA and %GC. Two very different patterns arose: strong selection likely affected the codon usage in Chlorophyta and mutation, and genetic drift appeared to be the major evolutionary driving forces for Streptophyta [10,11].

These patterns in Streptophyta mtDNA, produced essentially by non-adaptive forces, depended on the effective population size, generation times and the differences between unicellularity and multicellularity, which was consistent with the previous findings in plastids [12]. If strong selection was excluded, the challenge was to determine which other evolutionary force can explain the %GC increases throughout Streptophyta. Other possible explanations proposed for the GC bias, apart from selection, were the mutational bias and GC-biased gene conversion (gBGC) [13–15]. The ribosomal DNA (rDNA) genes are among the most conserved sequences in the three genetic compartments and have very specific evolutionary dynamics, including their long-term high recombination rate due to a concerted evolution.

They are also some of the most widely available sequences from eukaryotes in genomic databases. Therefore, they were considered useful genetic markers when analyzing the %GC variation throughout the entire genome. Based on the rDNA polymorphism data, the variation in nuclear rDNA %GC throughout the phylogenetic trees of angiosperms and vertebrates was observed [15] with a strong SNP (single nucleotide polymorphism) excess, for which either G or C was the majority allele. This was inconsistent with the mutational bias hypothesis and supported the GC-biased gene conversion (gBGC)/selection-driven evolution hypothesis.

The most outstanding aspect of this biased gene conversion was its impact on %GC, which affect the functional components of the genome and impeded natural selection (the Achilles' heel hypothesis) [14]. The gBGC appeared to play a significant role in the evolution of the genetic systems (e.g., sexual reproduction and recombination, inbreeding avoidance mechanisms, and ploidy cycles) and the development of the senescence and degeneration of non-recombining regions [16]. Despite the importance of this evolutionary force, the phylogenetic gBGC distribution in Streptophyta was not yet studied in association with the transition from water to land at the three genetic compartments.

Is Archaeplastida mitochondrial %GC related to the GL, %NC, NRS, gene number (GN) or coding sequences? If so, how are they linked and what role does this play in different lineages? Is there heterogeneity in the base composition through the streptophytes phylogeny? How is the %GC distributed in the three genetic compartments throughout the Streptophyta tree? Is the %GC increase concomitant for the three genome compartments? To answer these questions, our aims were the following: (i) To analyze the genomic variables with phylogenetically independent linear models in order to depict the evolutionary patterns of the mitochondrial genome in Archaeplastida. (ii) To examine the heterogeneity in base composition among the branches of the Streptophyta tree, using

non-homogeneous models of sequence evolution, taking into account the phylogenetic relationships. (iii) To implement a reconstruction of the ancestral GC content and GC* (equilibrium GC frequency) [17] in the three genetic compartments throughout the Streptophyta phylogenetic tree to compare their evolution.

2. Results

2.1. Genome Features

Across Archaeplastida evolution (Figure 1), taking all clades and studied variables into account (Supplementary Table S1), the only significant relationship found was the one between the NRS and GL, with both logarithms transformed for linearity (Supplementary Materials Figure S1A). These two variables were also linearly related to %NC; however, these relationships became unclear when the Streptophyta lineage was removed from the analyses, and became more significant when the Chlorophyta lineage was the one excluded (Supplementary Materials Figure S1B and C).

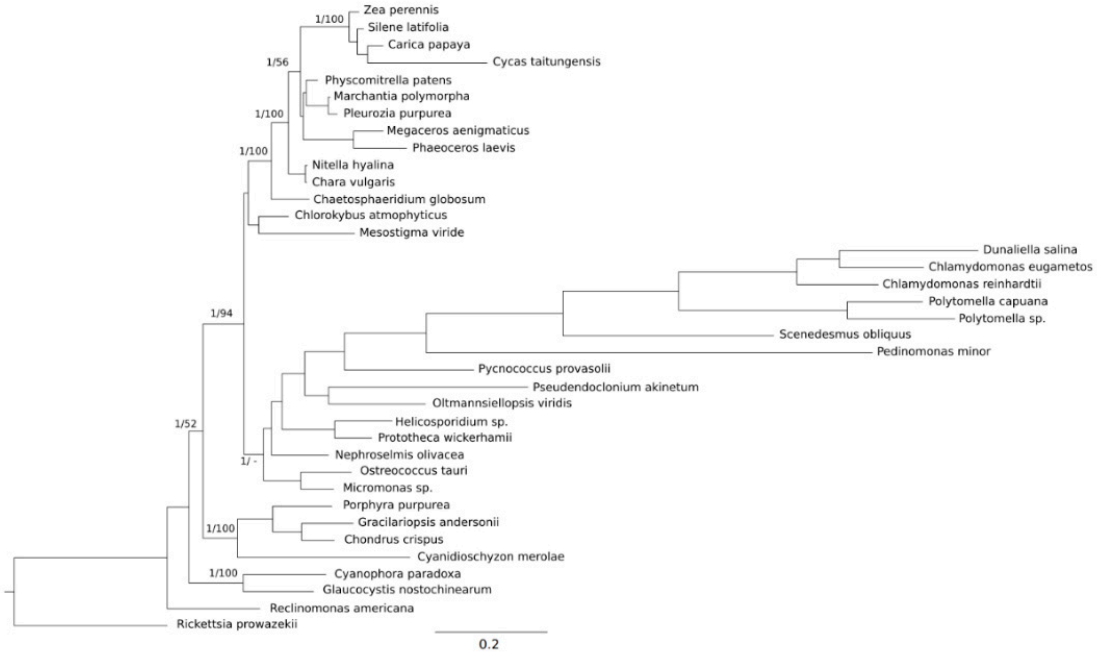

Figure 1. The phylogenetic relationships of the main clades of Archaeplastida, determined by six concatenated mtDNA-encoded proteins (cob, cox1, nad1, nad4, nad5, nad6), shared between the 37 studied genomes. The phylogenetic analyses were summarized by a Bayesian inference (BI) (cpREV model) consensus tree, with branch support from both BI and maximum likelihood (ML) analyses (Bayesian posterior probabilities/ML bootstrap). The prokaryote *Rickettsia prowazekii* and the protozoan excavate *Reclinomonas americana* were used as the outgroups. The scale bar represents the estimated number of amino acid substitutions per site.

Excluding the Chlorophyta lineage, the analyses determined an evident increase of the GL, %GC, %NC and NRS, across the Streptophyta lineage. The %GC exhibited a strong significant linear relationship with the %NC and log (NRS) (Figure 2A and B; Supplementary Materials Figure S1D and E). The effect of this last variable was mediated by the %NC (Figure 2C), and therefore, it lost its significance when the %NC effect was accounted for. In other words, the NRS affected the %NC and this explained 46% of the variance for the %GC. The GL was not directly related to the %GC, however it was clearly affected by the %NC, which explains 82 % of its variance (Figure 2D). The effect of the NRS over the GL was also mediated by the %NC. On the other hand, when the Streptophyta clade

was excluded from the analyses, the GL appeared to be related directly to the number of protein-coding genes (NPG) (Supplementary Materials Figure S1F).

Meanwhile, across Archaeplastida, the number of genes ranged greatly, both with lineages and between them. The main pattern was the gene number reduction in the Chlorophyceae family, along with an extreme reduction in the mitochondrial genome. The genes maintained in most species were those for tRNA, rRNA and ribosomal proteins and those involved in respiration and oxidative phosphorylation (Supplementary Materials Figure S2 and Table S2). Of these, only six genes were preserved in the mtDNA in all the studied species (cob, cox1, nad1, nad4, nad5 and nad6). Therefore, mitochondria lost most of the original bacterial genes and conserved only those associated with their principal cellular functions, respiration and oxidative phosphorylation, and those involved in the genetic machinery (rRNA and tRNA), but not in all cases. Two hornworts (Nothoceros aenigmaticus and Phaeoceros laevis) lost most of the ribosomal protein-coding genes [18,19], as did *Selaginella*, which also lost all tRNA genes [20].

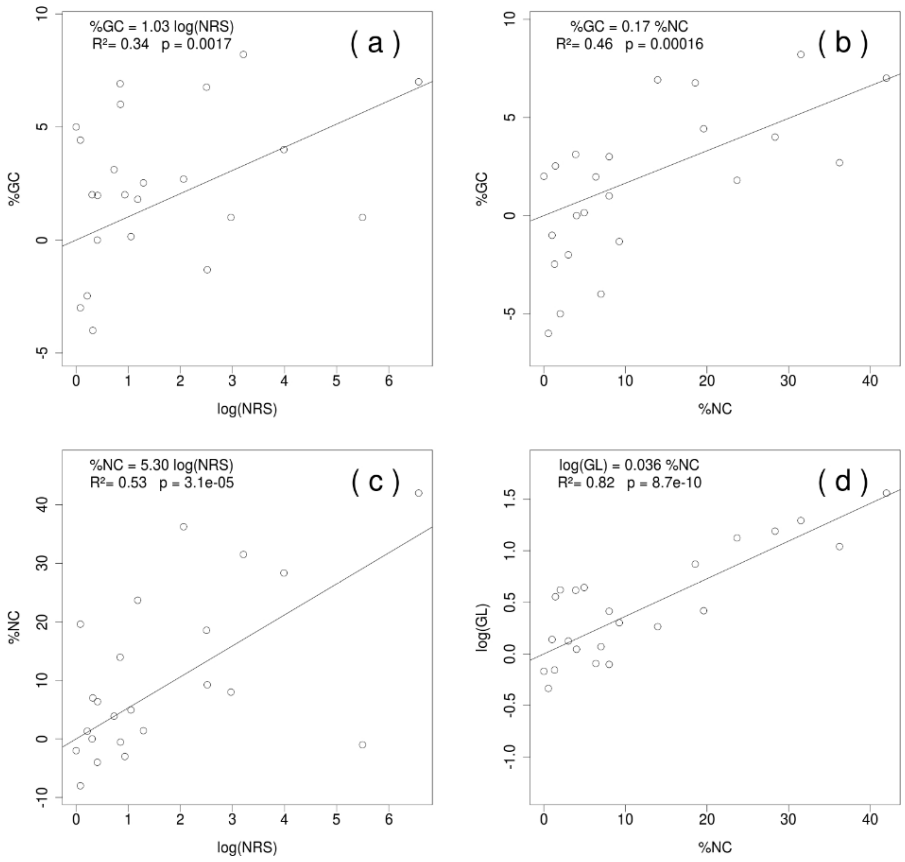

Figure 2. Phylogenetically independent contrasts (crunch) for the following relationships: Guanine-Cytosine (GC) content (%GC) with the number of repeated sequences (NRS) log transformed (**A**), and with the non-coding sequences (%NC) (**B**); between the %NC and NRS (**C**); and genome length (GL) with the %NC (**D**). All models were done excluding the species of Chlorophyta. For the complete set of contrasts, see the complementary material (Supplementary Materials Figure S1).

In addition, Chlorophyceae did not maintain any genes for ribosomal proteins, and the tRNAs were reduced considerably (Scenedesmus obliquus retained more tRNA genes but lost all ribosomal genes: rRNA and ribosomal protein-coding genes).

2.2. Heterogeneity in the Base Composition of Ribosomal Subunits in Streptophyta and the Reconstruction of the Ancestral %GC and GC Content*

The likelihood ratio test of non-homogeneous models showed the heterogeneity in the base composition among the clades studied. The non-homogeneous "terminal clades" hierarchical model fit the sequence heterogeneity better for the mtLSU, the cpLSU and the cpSSU (Table 1). Through the evolution of the streptophytes, from unicellular algae to angiosperms, through charophytes, mosses, ferns, and other lineages, the mtSSU increased its %GC. However, this was not in a progressive manner, but was variable according to the clades (Figure 3).

Table 1. Non-homogeneous models of substitutions.

Hierarchical Models	**Mitochondrion**							
	Large Ribosomal Subunit (LSU)				**Small Ribosomal Subunit (SSU)**			
	−InL	**Dev.**	**Df.**	***P*-Value**	**−InL**	**Dev.**	**Df.**	***P*-Value**
Homogeneous	22486.50				14603.13			
NH-Model_M1[(1)]	22438.91	95.18	17	6.88×10^{-13}	14556.56	93.14	15	2.57×10^{-13}
NH-Model_M2[(2)]	22436.54	4.73	1	0.0296	14556.50	0.12	1	0.7234
NH-Terminal clades (as Figure 3)	22425.82	21.43	3	8.56×10^{-5}	14548.21	16.58	3	0.0009
NH-One GC* per branch	22327.36	196.92	176	0.1337	14415.88	264.66	178	$2.66x\ 10^{-5}$
Nucleous								
					Small Ribosomal Subunit (SSU)			
					−InL	**Dev.**	**Df.**	***P*-Value**
Homogeneous					14023.18			
NH-Model_M1[(1)]					13984.94	76.47	19	7.47×10^{-9}
NH-Model_M2[(2)]					13970.40	29.08	1	6.94×10^{-8}
NH-Terminal clades (as Figure 3)					13967.78	5.23	3	0.1555
NH-One GC* per branch					13878.57	178.42	174	0.3933
Chloroplast								
	Large Ribosomal Subunit (LSU)				**Small Ribosomal Subunit (SSU)**			
	−InL	**Dev.**	**Df.**	***P*-Value**	**−InL**	**Dev.**	**Df.**	***P*-Value**
Homogeneous	23520.96				9895.45			
NH-Model_M1[(1)]	23324.90	392.13	19	0	9816.46	157.97	19	0
NH-Model_M2[(2)]	23323.04	3.71	1	0.0540	9816.37	0.17	1	0.6795
NH-Terminal clades (as Figure 3)	23287.39	71.31	3	2.22×10^{-15}	9808.42	15.90	3	0.0012
NH-One GC* per branch	23083.32	408.13	174	0	9695.43	225.98	174	0.0049

[(1)] NH-Model_M1: Non-homogeneous model which brings together to various clades; Angiosperms-Gymnosperms (seed plants) + Monilophyta-Lycopodiophyta + Bryophyta-Marchantiophyta + Anthocerotophyta + Charophyceae + Coleochaetophyceae + Zygnematophyceae + Klebsormidiophyceae + Chlorokybophyceae + Mesostigma. [(2)] NH-Model_M2: Like the model M1 but splitting the clades Marchantiophyta and Bryophyta. Note: -lnL: log likelihood; Dev.: residual deviance; df (degrees of freedom): residual degrees of freedom.

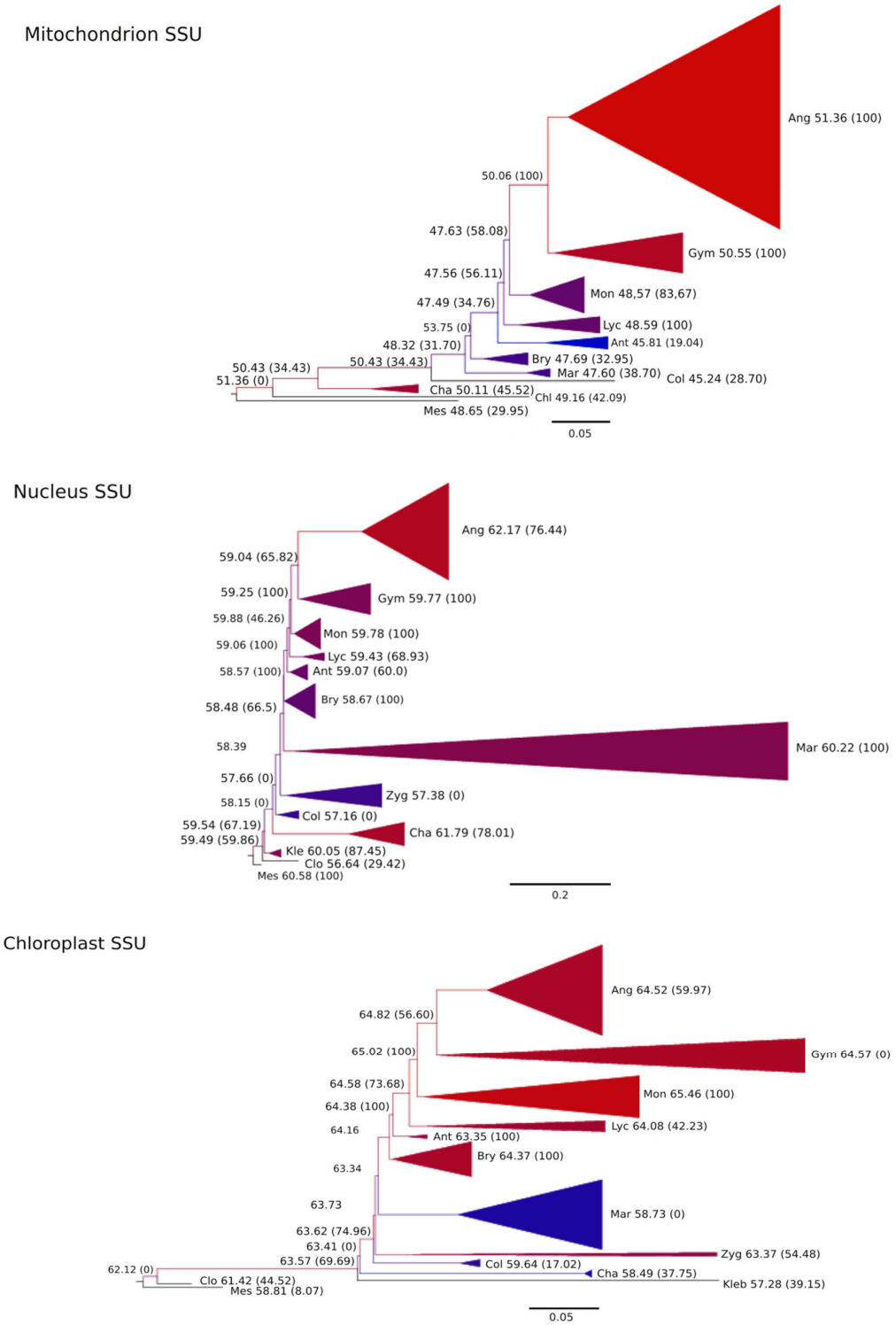

Figure 3. The evolution of the GC content and GC* across the phylogeny of Streptophyta: (A) Mitochondrial ribosomal Small Subunit (SSU); (B) Nuclear ribosomal SSU; (C) Plastidial ribosomal SSU. The values correspond to the ancestral GC content at the nodes, or the GC* (equilibrium GC content) in parentheses. The colors in the terminal branches represent the average GC content (blue: lowest GC content; red: highest GC content). The color scale is relative to the data set in each tree and is not directly comparable between them. The list of species and GC content at the terminal branches of each ribosomal subunit is available in Supplementary Table S3.

This process exhibited a clear pattern for the mtSSU, with differences among the seed plants, ferns, and club mosses, as well as the other lineages. The ribosomal nSSU resembled this augmentation pattern, but it was much less clear for the cpSSU. In all cases, the mtSSU contained noticeably less %GC than the cpSSU and the nSSU. The nSSU %GC ranged from 57.16% in the Coleochaetophyceae to 62.17% in the Angiosperms and, regarding the mtSSU, from 45.24% in the Coleochaetophyceae to 51.36% in the Angiosperms (a 6% increase). The cpSSU also demonstrated a %GC augmentation pattern, ranging from the Klebsormidiophyta at 57.28% to 65.46% in the Monilophyta (an 8.3% increase), with a high %GC for the Zygnematophyceae (63.37%) and a low %GC for the Charophyceae (58.49%) and Coleochaetophyceae (59.64%). Although not concurrently, these patterns are similar for the LSU, for both organelles studied (Supplementary Materials Figure S3).

A notable result was the low GC* in the mitochondrial Anthocerotophyta ribosomal subunits (19.04%). This very low value seemed to be unrelated to any of the biological or genomic characteristics studied; thus, we had no explanation for this phenomenon, despite the peculiarly high amount of editing in this phylum. The large %GC and GC* in the Charophyceae nSSU were also surprising. This pattern was maintained when the analyses were repeated without invariant sites.

Furthermore, we thoroughly looked for concomitancy of the patterns of the changes in the %GC, between the genetic compartments associated with the species and clades. We did not find any clear relation apart from the one between the large and small plastidial subunits, and only when coincident species were used. Therefore, we can ensure that although being present in the three genetic compartments, the increase in %GC followed a distinct pattern for each one, not concurrently through the lineages.

3. Discussion

3.1. Archaeplastida Mitochondrial Trends

The Archaeplastida mitochondrial genomes demonstrated evidence that the evolution of these genomes followed differentiated paths between the major lineages, as the evolutionary process is unpredictable due to many chance events. In the framework of population genetics theory, the mutational-hazard hypothesis predicted a favorable environment for the proliferation of non-coding mtDNA with a low product between the effective gene number per locus in the population (Ng) and the mutation rate (μ). The small Ngμ value in the mtDNA of green plants ($Ng\mu \ll 1$) compared to animals ($Ng\mu \gg 1$) intensified the genetic drift, making it easier for alleles with high mutation rates to behave neutrally, and thereby encouraging their fixation in the population [21]. This phenomenon could explain the difference between the small size and high mutation rate of animal mitochondrial genomes and the dramatic accumulation of non-coding mtDNA and the low mutation rate through the evolution of green plants.

In Streptophyta, an increase in the NRS incremented the %NC, which in turn facilitated the recombination and consequently raised the %GC through biased gene conversion. The best explanations for this observed gain in the GL and %GC appeared to be the mutational-hazard and the GC-biased gene conversion (gBGC) hypotheses, respectively. The promoted recombination was followed by an increase in %GC through biased gene conversion [13,21].

These non-adaptive forces occurring along the transition from water to land represented an exaptive genomic platform that explained some of the observed directional trends in the evolution of the land plant genomes: the recombination extension and the expansion of the genomic regulatory areas provoked a progressive accumulation of certain protein families correlated with the cell type number, which was essentially caused by gene duplication, concomitantly with the increase in organism complexity [22].

The chlorophytes are generally unicellular and sometimes parasites. They exhibit a GL reduction when the gene number decreases. By contrast, the Streptophyta evolution led to multicellular land plants, with a combination of striking features: organismal complexity together with a dramatic increase of the GL and %GC on the mitochondria [23].

The obtained models from our results were very consistent despite the relatively low taxon sampling. These observations supported the hypothesis that the changes in these variables were caused by the same factors. Thus, genome enlargement could provide a favorable environment for an increase in the recombination rate; therefore, a greater number of mismatches (that should be repaired by specific enzymes) would be produced, raising the probability for the incorrect nucleotides being replaced by G or C via gene conversion bias. Consequently, the %GC could be used as a fingerprint for the amount of recombination in these genomes. These mtDNA recombination mechanisms (surveillance and repair recombination-induced DNA damage, including mismatch repair) probably needed the selective pressure for efficient repair to be relaxed, raising the mutation rate [13]. The mitochondrial complexity augmentation throughout evolution did not occur in the mtDNA of multicellular animal lineages, although it did in their nuclear genome [24].

In Archaeplastida, the three genetic compartments must interact in coordination; consequently, a strong interactive adaptation between the organelles and the nucleus was necessary during evolution to multicellularity [25]. Hence, the three genomes increased their biological complexity in land plants, which permitted them to perform different functions in different tissues. Generally, species with a high %GC cpDNA also have %GC mtDNA but there are many exceptions [26,27]. In fact, evolutionary factors acting on the organism are the same in all three genetic compartments and consequently the differences indicated their variate idiosyncrasy and origin.

3.2. The Distribution of GC Content and GC in the Three Genetic Compartments for Streptophyta*

Basal Streptophyta algae evidenced a different pattern for %GC in their genetic compartments than Charophyceae and the terrestrial lineages. The %GC increased, especially from early land plants to angiosperms, yet there were intriguing exceptions to this pattern, such as the Charophyceae high %GC and GC* on the nSSU and mtSSU and the low value on cpSSU.

In the most basal streptophytes algae, the disparity in the %GC between the ribosomal subunits of the three genetic compartments resulted from strong selection, rather than from the gBGC, following a similar pattern to that observed in Chlorophyta [28]. Some aspects of the population biology of these lineages may have also been indirectly responsible for the %GC, because, as mentioned above, the population size and mutation rate determined that GL. *Chlamydomonas* is a Chlorophyceae algae living in freshwater habitats, with a similar ecology to Mesostigma; therefore, these genera are expected to have similar, very large populations.

The Streptophyta algae are multicellular (except for the Desmidiales), though very small (except for the Charophyceae), and, consequently, they may maintain large populations. In contrast, the Charophyceae have a large body size, with few individuals living in ponds or lagoons, and, hence, a very small effective population size, although they are multinucleated. Terrestrial plants also have small populations compared to those of the basal Streptophyta algae. In fact, Harholt et al. [29] hypothesized that streptophyte algae lived on land before the emergence of embryophytes, which was verified, recently, by Wang et al. [3], in the common ancestor of *Mesostigma viride* and *Chlorokybus atmophyticus*, where the development of traits reflected adaptations to a subaerial/terrestrial habitat.

The reasons for the differences in the recombinations and the gBGC among genomes, species and lineages remains unknown. Nevertheless, in the Charophyceae, these high %GC and GC* in the nucleus can be explained by the large C value (2C nuclear DNA content) (10–50 pg), which was higher than for the rest of the streptophytes algae (0.7–5 pg) and all mosses studied (0.9–5 pg) [30].

4. Materials and Methods

4.1. Archaeplastida Mitochondrial Genome-Wide Characteristics

All 35 Archaeplastida mitochondrial genomes present in the NCBI database were selected. Two outgroups were also chosen for their unique characteristics. The protozoan *Reclinomonas americana* (Excavata) was the shortest mitochondrial genome from eukaryotes and probably the one that best reflected the ancestral state [31]. Additionally, the parasite *Rickettsia prowazekii* (α-proteobacteria) is one of the closer derivatives from the eubacterial organisms, where the mitochondrial organelles

originated [32]. The samples represented the major Archaeplastida clades, including Glaucophyta (n = 2), Rhodophyta (n = 4), Prasinophyta (n = 4), basal Chlorophyta (n = 5), Chlorophyceae (n = 6), basal Streptophyta (n = 3), Charophyceae (n = 2), including *Nitella hyalina* (GenBank database code JF810595), basal Embryophyta (n = 5), and Spermatophyta (n = 4). Certain features of the genome, including the %GC, were extracted from the databases at NCBI, using Artemis software [33]. The following variables were also obtained from all species: GL, %GC, %NC, NPG, NRS and RSL. The complete list of species, accession numbers and genome characteristics are available in the supplementary materials (Supplementary Materials Table S1).

To find the repeated sequences, the complete mtDNA sequences (forward, reverse, complement and reverse complement) from selected species were analyzed with REPuter [34]. The minimal repeat size was limited to 20 nucleotides (nt) for general analyses, 50 nt for less repetitive mitochondrial sequences and 100 nt in more repetitive sequences, when using the largest repeats in each genome. The %GC within repetitions was determined using Emboss [35].

Linear model analyses were used to determine the shape and significance of the relationship between each pair of variables. These linear models were implemented in three different ways to test for consistency: The classic linear regression and two different phylogenetically independent contrasts: "the Phylogenetic Generalized Least Squares (PGLS) method and "crunch" method, both from the "caper" package in R [36]. PGLS was a powerful method to estimate the adaptive optima using continuous data [37]. This method assumed that the analyzed trait evolved by Brownian motion and thus trait covariance between any pair of taxa decreased linearly with time (branch length) since their divergence. The methods were originally provided in the programs CAIC [38] and MacroCAIC [39]. Both programs calculated phylogenetically independent contrasts in a set of variables and then used linear models of those contrasts to test for evolutionary relationships. All contrast model functions enforced regression through the origin. Mediation tests were also performed to determine whether the relationships between pairs of variables were mediated by a third variable.

All these analyses were carried out on three different data sets: (i) all the studied species without the outgroups, (ii) eliminating the Chlorophyta, and (iii) eliminating the Streptophyta. These two lineages evolved by different evolutionary paths, and therefore represented different non-comparable patterns.

The assumptions of normality and homoscedasticity of the residuals were evaluated to verify the appropriateness of the linear modeling. A Bonferroni correction was carried out to correct the results from all these multiple tests and therefore, linear relationships were considered significant only when $p < 0.005$.

The phylogenetic tree used for the independent contrasts was built from the only six protein-coding genes that were shared by the 37 species studied (cob, cox-1, nad1, nad4, nad5 and nad6). These genes sequences were translated into amino acids with TranslatorX [40], aligned with Muscle [41] and trimmed with GBLOCKs [42] with the default parameters, resulting in a concatenated matrix with 1725 amino acids. The sequences matrix for each gene was subjected to ProtTest to find the best fit evolutionary model [43]. In order to test the phylogenetic signal TREE-PUZZLE was used [44]. For the maximum-likelihood (ML) analyses, the concatenated protein matrix was analyzed with RAxML v. 7.2.8 [45] using the WAG model [46] and a bootstrap analysis with 1000 replicates.

Bayesian analyses were implemented with Mr. Bayes V.2.1.0 [47]. The concatenated protein matrix was analyzed using three model partitions: two fixed cpREV models [48] of amino acid substitution, with inv-gamma and gamma distributions of rates, and the third model with a fixed Jones model with gamma distribution of rates. The analyses, in all cases, consisted of three million generations, four independent runs and four Markov chains. The trees were sampled every 1000 generations; stationarity was assessed by examining the standard deviation of the split frequencies and by plotting the –ln L per generation using Tracer v1.4 [49], and the trees were generated before the stationarity was discarded. Further Bayesian analysis with a fixed-cpREV model for the six coincident mitochondrial protein-coding genes was consistent with the tree obtained with ML and

proved to be the best-fitting tree to most accepted phylogenies. Therefore, this tree was utilized for the statistical phylogenetically independent analyses.

4.2. Analyses of Heterogeneity in Base Composition

Mitochondrial ribosomal small and large subunits (mtSSU and mtLSU), chloroplast ribosomal subunits (cpSSU and cpLSU), and nuclear ribosomal small subunit (nSSU) sequences were downloaded from the SILVA database [50]. A hundred species for each ribosomal subunit were selected in order to cover all available lineages across the Streptophyta tree, discarding short, incomplete or highly gapped sequences. The species names for each subunit and genetic compartment are available in the supplementary material (Supplementary Materials Table S3).

Zygnematophyceae were not represented in the mitochondrial subunits, nor were Klebsormidiophyceae in the mtSSU, as there were no available sequences in the databases. Homologous rRNAs for each subunit and genetic compartment were aligned using ClustalW [51]. Escobar et al. [15] eliminated the hypermutable CpG sites to prevent slippage in the angiosperm %GC calculation. However, as Archaeplastida lineages diverged very early, hypermutable CpG site detection cannot be performed in a reliable way, so this step was omitted from the analyses [52]. The length of the sequences used had approximated sizes of 1289 (mtSSU), 1384 (mtLSU), 1272 (nSSU), 1162 (cpSSU) and 2155 (cpLSU) base pairs.

Phylogenetic trees were inferred with PhyML.3.0 [53] in the five RNA markers used. The models of sequence evolution were obtained with the program JModeltest2 [54] using the general time reversible model (GTR) + I + G, with four categories for the gamma distribution, parsimony starting trees and SPR (sub-tree pruning and regrafting) branch swapping. In general, terminal clades were grouped consistently with the most current Streptophyta phylogeny, although deep clades were less resolved. Therefore, some branches of the trees resulting from the above analyses were relocated with Baobab software [55], in order to adjust them to the accepted phylogenies [56]. We thensubsequently re-optimize the branch length with PhyML. The ML analyses of non-homogeneous models presented here were implemented with the modified trees. However, analyses with unmodified trees were performed to test for robustness.

The heterogeneity in %GC was tested with four non-homogeneous models of sequence evolution. These models were fitted with the software BppML [57] and NHML [58], which utilize an ML approach that is not stationary (ancestral and current %GC may differ) and with no homogeneity (the branches may have different %GC) in the base composition across the phylogeny. Thus, the %GC at the nodes and the GC* at branches of the phylogenetic tree were estimated. These hierarchical models were fitted to test whether the branches underwent similar evolution in their base composition or not, using the fixed trees from PhyML. Nucleotide substitution models based on Galtier and Gouy [58] were implemented for all BppML analyses, using four gamma categories and two parameters: theta (GC*) and kappa (Ts/Tv).

The hierarchical models used were: 1) a homogeneous model; 2) a non-homogeneous model (M1), Angiosperms-Gymnosperms (seed plants) + Monilophyta-Lycopodiophyta + Bryophyta-Marchantiophyta + Anthocerotophyta + Charophyceae + Coleochaetophyceae + Zygnematophyceae + Klebsormidiophyceae + Chlorokybophyceae + Mesostigma; 3) a non-homogeneous model (M2), which was the same as above but splitting Marchantiophyta and Bryophyta; 4) a non-homogeneous model of terminal clades; and 5) a non-homogeneous model. One was used per branch, where each branch had its own %GC and GC*. The likelihood ratio test (LRT) was used to assess whether more complex nested models provided a significantly improved fit compared with simpler models.

*4.3. Streptophyta Ribosomal GC Content and GC**

The %GC and GC* were estimated for all species and nodes across the Streptophyta phylogenetic tree, using the same method that Escobar et al. [15] used. Therefore, the GC* was defined as:

$$GC = \frac{AT \rightarrow GC}{AT \rightarrow GC + GC \rightarrow AT} \quad (1)$$

where AT→GC refers to the substitution rate from A or T to G or C bases and GC→AT holds for the inverse. The GC* was considered a more appropriate estimator for evolutionary dynamics than the %GC, because GC* reflects the relative contribution of changes from AT to GC independently from the total number of mutations [59]. NHML software [58] was used to implement a ML approach for the non-stationary probabilistic and non-homogeneous model. The objective set was to reconstruct the ancestral %GC and GC* distributions optimizing parameters on the model "terminal clades" over the trees obtained from BppML, for each one of the five ribosomal subunits. The analyses were performed with and without invariable sites to check the consistency. Sequence gaps were removed in all cases. The results from these analyses were two phylogenetic trees for each of the ribosomal subunits analyzed, with pseudo-bootstrap values for %GC and GC*. The GC* is the theta parameter estimated in this ML framework [58].

4.4. Concomitant Evolution between Genetic Compartments and Clades

With the aforementioned trees, phylogenetically independent contrasts (PGLS and "crunch") of %GC were carried out to check for concomitant evolution. The following contrasts were made: between the small and the large subunits for each genetic compartment; between the large subunits of the mitochondrion and plastid; between the small subunits of the three different genetic compartments. These tests were performed in two different ways: in the first instance, using the %GC in the shared species between subunits pairs (Supplementary Materials, Table S3), and second, using the node values (NHML) in the Streptophyta clades, despite the species not being the same. In both cases, each analysis was implemented with three different tree topologies: the phylogenetic tree of the two subunits studied and the tree with a fixed topology (all branch lengths equal to one). These analyses were also conducted with and without invariant sites and with the Bonferroni corrected for multiple (six) comparisons, considering significant differences only when $p < 0.008$.

Supplementary Materials: The following are available online at www.mdpi.com/2223-7747/9/3/358/s1, Table S1: Genomic variables studied: Species, Genbank accession number, clade, genome length (GL), GC content (%GC), non-coding DNA (%NC), gene number (GN), number of protein-coding genes (NPG), number of repeated sequences (NRS) and repeated sequences total length (RSL), Table S2: Presence of mitochondrial genes (1). Number of genes/species and number of studied species/gene are given, Table S3: GC content in the ribosomal subunits (RSU) from the 100 species of Streptophyta used in the analysis at each subunit. List of species used to estimate the phylogeny of each ribosomal subunit, Figure S1: Linear regression (1.dashed lines), Phylogenetic Generalized Least Squares (PGLS) method (1. continuous lines) and "crunch" method (2) for the relationships between genomic variables, with all species (black), excluding the species of Chlorophyta (red) and excluding the species of Streptophyta (green): A) The effect of the log transformed number of repeated sequences (NRS), on non-coding genome (%NC); B) Effect of %NC on log transformed Genome length (GL); C) Effect of log(NRS) on GC content (%GC); D) Effect of %NC on %GC; E) Effect of the number of protein-coding genes (NPG) on the log(GL), Figure S2: Mitochondrial gene classes across Archaeplastida. Average gene number for each function in each clade, Figure S3: Evolution of GC content and GC* across the phylogeny of Streptophyta: A) Mitochondrial ribosomal LSU; B) Plastidial ribosomal LSU. Values correspond to ancestral GC content at nodes, or GC* (equilibrium GC content) in parentheses. Colors in terminal branches represent average GC content (blue: lowest GC content; red: highest GC content). The color scale is relative to the data set in each tree and is not directly comparale between them. List of species and GC content at terminal branches of each ribosomal subunit is available in Supplementary Table S3.

Author Contributions: J.P.-M. and L.P. conceived, designed and supervised the project. J.P.-M., D.L.-G., C.G.-B., L.P., A.G, C.S. and J.M.S. provided resources and materials, developed the protocol and analyzed samples and data. J.P.-M. and J.M.S. wrote the paper.

Funding: This research was funded by the European Commission (Environment – LIFE Programme) project for the Comunidad Valenciana (Spain), LIFE05 NAT/E/000060.

Acknowledgments: We thank Amparo Latorre, Rosario Gil, and Juan Antonio Delgado for their helpful comments on the article. The newly reported sequence for the complete mitochondrial genome from the Charophyceae *Nitella hyalina* has the GenBank accession number JF810595.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

%NC	non-coding DNA
cpLSU	chloroplast ribosomal large subunits
cpSSU	chloroplast ribosomal small subunits
gBGC	GC-biased gene conversion
GC*	GC frequency
GL	genome lengths
GN	gene number
mtDNA	mitochondrial genome
mtLSU	mitochondrial ribosomal large subunits
mtSSU	mitochondrial ribosomal small subunits
NPG	number of protein-coding genes
NRS	number of repeated sequences
nSSU	nuclear ribosomal small subunit
nt	nucleotides
PGLS	Phylogenetic Generalized Least Squares
rDNA	ribosomal DNA
RSL	repeated sequences total length

References

1. Melton, J.T., III; Leliaert, F.; Tronholm, A.; Lopez-Bautista, J.M. The complete chloroplast and mitochondrial genomes of the green macroalga *Ulva* sp. UNA00071828 (Ulvophyceae, Chlorophyta). *PLoS ONE* **2015**, *10,* e0121020.
2. Green, B.R. Chloroplast genomes of photosynthetic eukaryotes. *Plant J.* **2011**, *66*, 34–44.
3. Wang, S.; Li, L.; Li, H.; Sahu, S.K.; Wang, H.; Xu, Y.; Xian, W.; Song, B.; Liang, H.; Cheng, S.; et al. Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat. Plants* **2020,** *6*, 95–106.
4. Kim, E.; Graham, L.E. EEF2 analysis challenges the monophyly of Archaeplastida and Chromalveolata. *PLoS ONE* **2008**, *3*, e2621.
5. Chan, C.X.; Gross, J.; Yoon, H.S.; Bhattacharya, D. Plastid origin and evolution: New models provide insights into old problems. *Plant Physiol.* **2011**, *155*, 1552–1560.
6. Adams, K.L.; Palmer, J.D. Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Mol. Phylogenetics Evol.* **2003**, *29*, 380–395.
7. Smith, D.R.; Crosby, K.; Lee, R.W. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biol. Evol.* **2011,** *3*, 365–371.
8. Smith, D.R.; Keeling, P.J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10177–10184.
9. Kern, R.; Facchinelli, F.; Delwiche, C.; Weber, A.P.; Bauwe, H.; Hagemann, M. Evolution of photorespiratory glycolate oxidase among Archaeplastida. *Plants* **2020**, *9*, 106.
10. Wang, B.; Liu, J.; Jin, L.; Feng, X.Y.; Chen, J.Q. Complex mutation and weak selection together determined the codon usage bias in bryophyte mitochondrial genomes. *J. Integr. Plant. Biol.* **2010**, *52*, 1100–1108.
11. Wang, B.; Yuan, J.; Liu, J.; Jin, L.; Chen, J.Q. Codon usage bias and determining forces in green plant mitochondrial genomes. *J. Integr. Plant. Biol.* **2011**, *53*, 324–334.
12. Mower, J.P.; Sloan, D.B.; Alverson, A.J. *Plant Mitochondrial Genome Diversity: The Genomics Revolution*; Plant Genome Diversity; Springer: Vienna, Austria, 2012; Volume 1, pp. 123–144.
13. Galtier, N.; Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* **2007**, *23*, 273–277.
14. Galtier, N.; Duret, L.; Glémin, S.; Ranwez, V. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* **2009**, *25*, 1–5.
15. Escobar, J.S.; Glémin, S.; Galtier, N. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Mol. Biol. Evol.* **2011**, *28*, 2561–2575.

16. Lassalle, F.; Périan, S.; Bataillon, T.; Nesme, X.; Duret, L.; Daubin, V. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. *PLoS Genet.* **2015**, *11*, e1004941.
17. Sueoka, N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* **1962**, *48*, 582–592.
18. Li, L.; Wang, B.; Liu, Y.; Qiu, Y.-L. The complete mitochondrial genome sequence of the hornwort *Megaceros aenigmaticus* shows a mixed mode of conservative yet dynamic evolution in early land plant mitochondrial genomes. *J. Mol. Evol.* **2009**, *68*, 665–678.
19. Xue, J.Y.; Liu, Y.; Li, L.; Wang, B.; Qiu, Y.L. The complete mitochondrial genome sequence of the hornwort *Phaeoceros laevis*: Retention of many ancient pseudogenes and conservative evolution of mitochondrial genomes in hornworts. *Curr. Genet.* **2010**, *56*, 53–61.
20. Hecht, J.; Grewe, F.; Knoop, V. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: The root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biol. Evol.* **2011**, *3*, 344–358.
21. Lynch, M.; Walsh, B. *The Origins of Genome Architecture*; Sinauer Associates: Sunderland, UK, 2007; Volume 98.
22. Vogel, C.; Chothia, C. Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2006**, *2*, e48.
23. Turmel, M.; Lemieux, C. Evolution of the plastid genome in green algae. In *Advances in Botanical Research*; Academic Press: New York, NY, USA, 2018; Volume 85, pp. 157–193.
24. Wallace, D.C. Bioenergetics, the origins of complexity, and the ascent of man. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 8947–8953.
25. Nishiyama, T.; Sakayama, H.; De Vries, J.; Buschmann, H.; Saint-Marcoux, D.; Ullrich, K.K.; Haas, F.B.; Vanderstraeten, L.; Becker, D.; Lang, D.; et al. The Chara genome: Secondary complexity and implications for plant terrestrialization. *Cell* **2018**, *174*, 448–464.
26. Chaitanya, K.V. Orgenellar Genome Analysis. In *Genome and Genomics*; Springer: Singapore, 2019; pp. 89–119.
27. Gitzendanner, M.A.; Soltis, P.S.; Wong, G.K.S.; Ruhfel, B.R.; Soltis, D.E. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *Am. J. Bot.* **2018**, *105*, 291–301.
28. Zheng, F.; Wang, B.; Shen, Z.; Wang, Z.; Wang, W.; Liu, H.; Wang, C.; Xin, M. The chloroplast genome sequence of the green macroalga Caulerpa okamurae (Ulvophyceae, Chlorophyta): Its structural features, organization and phylogenetic analysis. *Mar. Genom.* **2020**, in press.
29. Harholt, J.; Moestrup, Ø.; Ulvskov, P. Why plants were terrestrial from the beginning. *Trends Plant Sci.* **2016**, *21*, 96–101.
30. Kapraun, D.F. Nuclear DNA content estimates in green algal lineages: Chlorophyta and Streptophyta. *Ann. Bot.* **2007**, *99*, 677–701.
31. Lang, B.F.; Burger, G.; O'Kelly, C.J.; Cedergren, R.; Golding, G.B.; Lemieux, C.; Sankoff, D.; Turmel, M.; Gray, M.W. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **1997**, *387*, 493–497.
32. Yang, D.; Oyaizu, Y.; Oyaizu, H.; Olsen, G.J.; Woese, C.R. Mitochondrial origins. *Proc. Natl. Acad. Sci. USA* **1985**, 82, 4443–4447.
33. Carver, T.; Harris, S.R.; Berriman, M.; Parkhill, J.; McQuillan, J.A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **2012**, *28*, 464–469.
34. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **2001**, *29*, 4633–4642.
35. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology open software suite. *Trends Genet.* **2000**, *16*, 276–277.
36. Orme, D.; Freckleton, R.; Thomas, G.; Petzoldt, T.; Fritz, S.; Isaac, N.; Thomas, G.; Petzoldt, T.; Pearse, W.; Fritz, S.; et al. Caper: Comparative Analyses of Phylogenetics and Evolution in R. R Package Version 0.5. Available online: http://CRAN.R-project.org/package=caper URL (accessed on 20 February 2020).
37. Butler, M.A.; King, A.A. Phylogenetic comparative analysis: A modelling approach for adaptive evolution. *Am. Nat.* **2004**, *164*, 683–695.
38. Purvis, A.; Rambaut, A. Comparative analysis by independent contrasts (CAIC): An Apple Macintosh application for analysing comparative data. *Comput. Appl. Biosci.* **1995**, *11*, 247–251.

39. Agapow, P.; Isaac, N.J.B. MacroCAIC: Revealing correlates of species richness by comparative analysis. *Divers. Distrib.* **2002**, 8, 41–43.
40. Abascal, F.; Zardoya, R.; Telford, M.J. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **2010**, *38*, 7–13.
41. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
42. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552.
43. Abascal, F.; Zardoya, R.; Posada, D. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* **2005**, *21*, 2104–2105.
44. Schmidt, H.A.; Strimmer, K.; Vingron, M.; von Haeseler, A. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **2002**, *18*, 502–504.
45. Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **2006**, *22*, 2688–2690.
46. Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699.
47. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542.
48. Adachi, J.; Waddell, P.J.; Martin, W.; Hasegawa, M. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* **2000**; *50*, 348-358.
49. Drummond, A.J.; Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **2007**, *7*, 214.
50. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2012**, *41*, D590–D596.
51. Thomson, 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.* **1994**, *22*, 4673–4680.
52. Arndt, P.F.; Petrov, D.A.; Hwa, T. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **2003**, *20*, 1887–1896.
53. Torres, M.; Oliveira da Silva, J. Parallel solution based on collective communication operations for phylogenetic bootstrapping in PhyML 3.0. In: *Brazilian Symposium on Bioinformatics*; Springer: Cham, Switzerland, 2018; pp. 133–145.
54. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **2012**, *9*, 772.
55. Dutheil, J.; Galtier, N. BAOBAB: A java editor for large phylogenetic trees. *Bioinformatics* **2002**, *18*, 892–893.
56. Timme, R.E.; Bachvaroff, T.R.; Delwiche, C.F. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE* **2012**, *7*, 1–8.
57. Durtheil, J.; Boussau, B. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* **2008**, *8*, 255.
58. Galtier, N.; Gouy, M. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **1998**, *15*, 871–879.
59. Duret, L.; Arndt, P.F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **2008**, *4*, e1000071.