


RESEARCH

Open Access

The economics of organellar gene loss and endosymbiotic gene transfer



Steven Kelly 

Correspondence: steven.kelly@plants.ox.ac.uk
Department of Plant Sciences,
University of Oxford, South Parks
Road, Oxford OX1 3RB, UK

Abstract

Background: The endosymbiosis of the bacterial progenitors of the mitochondrion and the chloroplast are landmark events in the evolution of life on Earth. While both organelles have retained substantial proteomic and biochemical complexity, this complexity is not reflected in the content of their genomes. Instead, the organellar genomes encode fewer than 5% of the genes found in living relatives of their ancestors. While many of the 95% of missing organellar genes have been discarded, others have been transferred to the host nuclear genome through a process known as endosymbiotic gene transfer.

Results: Here, we demonstrate that the difference in the per-cell copy number of the organellar and nuclear genomes presents an energetic incentive to the cell to either delete organellar genes or transfer them to the nuclear genome. We show that, for the majority of transferred organellar genes, the energy saved by nuclear transfer exceeds the costs incurred from importing the encoded protein into the organelle where it can provide its function. Finally, we show that the net energy saved by endosymbiotic gene transfer can constitute an appreciable proportion of total cellular energy budgets and is therefore sufficient to impart a selectable advantage to the cell.

Conclusion: Thus, reduced cellular cost and improved energy efficiency likely played a role in the reductive evolution of mitochondrial and chloroplast genomes and the transfer of organellar genes to the nuclear genome.

Keywords: Endosymbiosis, Gene loss endosymbiotic gene transfer, Mitochondrion, Chloroplast, Organellar genome

Background

Endosymbiosis has underpinned two of the most important innovations in the history of life on Earth [1, 2]. The endosymbiosis of the alphaproteobacterium that became the mitochondrion led to the emergence and radiation of the eukaryotes [3–5], and the endosymbiosis of the cyanobacterium that became the chloroplast first enabled oxygenic photosynthesis in eukaryotes [6, 7]. The function and evolution of both organelles are inextricably linked with energy metabolism and the evolution of the



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

eukaryotic cell [4, 8–14], and together they have given rise to the multicellular organisms that constitute the largest fraction of the biomass of the biosphere [15].

Following the endosymbioses of the bacterial progenitors of the mitochondrion and the chloroplast, there was a dramatic reduction in the gene content of the endosymbiont genomes such that they harbor fewer than 5% of the genes found in their free-living bacterial relatives [16–18]. While many of the original endosymbiont genes have been lost [19–22], others have been transferred to the host nuclear genome and their products imported back into the organelle where they function—a process known as endosymbiotic gene transfer [23–27]. For example, the mitochondria of humans [28] and chloroplasts of plants [29] each contain more than 1000 proteins, yet their genomes encode fewer than 100 genes. Thus, the reduced gene content of organelles is not representative of their molecular, proteomic, or biochemical complexity.

The process of gene loss and endosymbiotic gene transfer is not unique to the evolution of chloroplasts and mitochondria but has also been observed concomitant with the endosymbioses of bacteria in insects [19, 30] and the endosymbiosis of the cyanobacterium that became the chromatophore in *Paulinella* [31–35]. In addition, it has been suggested that lateral gene transfers from diverse bacteria into the host nuclear genome may have contributed to the process of organellar genome reduction in a manner that functionally recapitulates endosymbiotic gene transfer, i.e., the endosymbiont gene becomes redundant when an orthologous or functionally equivalent gene from another species is transferred to the nuclear genome [36]. Similarly, the presence of a redundant copy of a gene in the nucleus that is only slightly expressed and minimally targeted provides an opportunity for recovery in case of gene loss from the organelle. Thus, endosymbiont genome reduction in the presence of functional compensation (by lateral and/or endosymbiotic gene transfer or pre-existing nuclear genes) is a recurring theme in the evolution of organellar and endosymbiont genomes.

Given the importance of endosymbiotic gene transfer (and functionally equivalent lateral complementation) to the evolution of eukaryotic genomes, several hypotheses have been proposed to explain why it occurs [37–41]. For example, it has been proposed that lateral and endosymbiotic gene transfer protects endosymbiont genes (and the biological functions they provide) from mutational hazard [20, 21, 41, 42] and that it enables endosymbiont genes that are otherwise trapped in a haploid genome to recombine and thus escape from Muller's ratchet [20, 21, 39, 43–45]. It has also been proposed that endosymbiotic gene transfer is an inevitable consequence of a constant stream of endosymbiont genes entering the nucleus [46–50], and that transfer to the nuclear genome allows the host cell to gain better control over the replication and function of the organelle [38] allowing better cellular network integration [33, 51]. However, mutation rates of organellar genes are often not higher than nuclear genes [20–22, 52–56], and therefore, effective mechanisms for protection against DNA damage in organelles must exist. Similarly, although there is evidence for the action of Muller's ratchet in mitochondria [44, 45], chloroplasts appear largely to escape this effect [52, 57] likely due to gene conversion [58], and thus, it does not fully explain why endosymbiotic gene transfer occurred in both lineages. Finally, the nature of the regulatory advantage for having genes reside in the nuclear genome is difficult to quantify, as bacterial gene expression regulation is no less effective than in eukaryotes, and many eukaryotes

utilize polycistronic regulation of gene expression [59–62]. Thus, it is unclear whether endosymbiotic gene transfer functions simply as rescue from processes that would otherwise lead to gene loss, or whether there may also be an advantage to the cell for transferring an endosymbiont gene to the nuclear genome.

Given the constant stream of genetic transfer to the nucleus, and the proposed reasons why these transfers may be advantageous, the question arises as to why organelles have retained any genetic material. To answer this question, several hypotheses have been put forward that suggest that there must be a selectable advantage for the retention of genes in organellar genomes. Foremost among these hypotheses is that the location of genes in organelles enables regulation of their expression by the redox state of the organelle [63–65]. In addition, analyses of thousands of organellar genomes led to the suggestion that other gene-intrinsic factors such as GC content or hydrophobicity of the gene product may also play a factor in providing an advantage for gene retention in organellar genomes [66, 67]. These collectively point to a role for natural selection in the retention of organellar genes in organellar genomes.

We hypothesized that an advantage for endosymbiotic gene transfer or retention of a gene in an organellar genome may arise from the difference in the cost to the cell of encoding a gene in the organellar and nuclear genome. This is because each eukaryotic cell typically contains multiple organelles and each organelle typically harbors multiple copies of the organellar genome [68, 69]. The number of organelles in a cell reflects the biochemical requirement of that cell for those organelles, and the high genome copy number per organelle has been proposed to provide protection against DNA damage [70] and to enable the organelle to achieve high protein abundance for genes encoded in the organellar genome [69]. Thus, while a diploid eukaryotic cell contains two copies of the nuclear genome, the same cell can contain hundreds to hundreds of thousands of copies of its organellar genomes [68, 69]. For example, endosymbiotic transfer of a 1000-bp gene from the mitochondrion to the nuclear genome in humans, yeast, or *Arabidopsis* would save 5,000,000 bp, 200,000 bp, or 100,000 bp of DNA per cell, respectively, and an analogous transfer from the chloroplast genome to the nuclear genome in *Arabidopsis* would save 1,500,000 bp of DNA per cell (see the “Methods” section for sources of genome copy numbers). As DNA costs energy and cellular resources to biosynthesize [71], we hypothesized that if the energy saved by transferring a gene from the organellar genome to the nuclear genome offset the cost of importing the encoded gene products (proteins) back into the organelle then this would provide a direct energetic advantage to the host cell for endosymbiotic gene transfer. Similarly, if a functionally equivalent gene from another species was laterally acquired by the nuclear genome, then there would be an analogous energetic advantage to the host cell to utilize the acquired gene and delete the organellar gene.

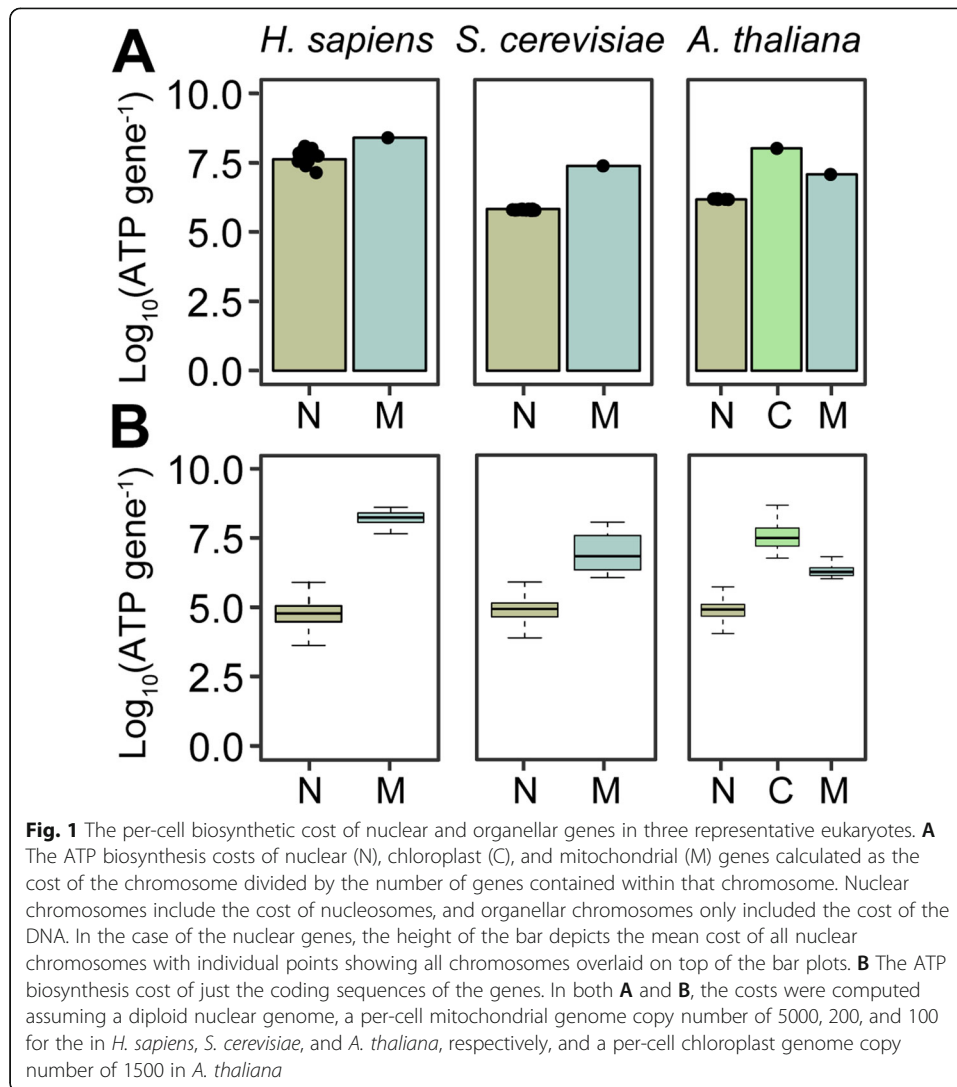
Here, we analyze the relative cost of DNA synthesis and protein import over a broad range of plausible parameter spaces for eukaryotic cells that encompasses total cell protein content, organellar fraction (i.e., the fraction of the total number of protein molecules in a cell that is contained within the organelle), organellar genome copy number, organellar protein abundance, organellar protein import cost, organellar protein import efficiency, cell life span, and protein turnover rate. Through this, we reveal that for the vast majority of plausible parameter space for eukaryotic cells, it is energetically favorable to the cell to transfer organellar genes to the nuclear genome and re-import the

proteins back to the organelle. We show that the interplay between per-cell organellar genome copy number and per-cell organellar protein abundance determines the magnitude of the energy saved such that it is only energy efficient for the cell to retain genes in the organellar genome if they encode proteins with very high abundance. Through analysis of the energy saved by endosymbiotic gene transfer in the context of total cellular energy budgets, we demonstrate that the net energetic advantage of endosymbiotic gene transfer is a significant proportion of total cell energy budgets and would thus confer a selectable energetic advantage to the cell. Collectively, these results reveal that enhanced energy efficiency has helped to shape the content and evolution of eukaryotic organellar and nuclear genomes.

Results

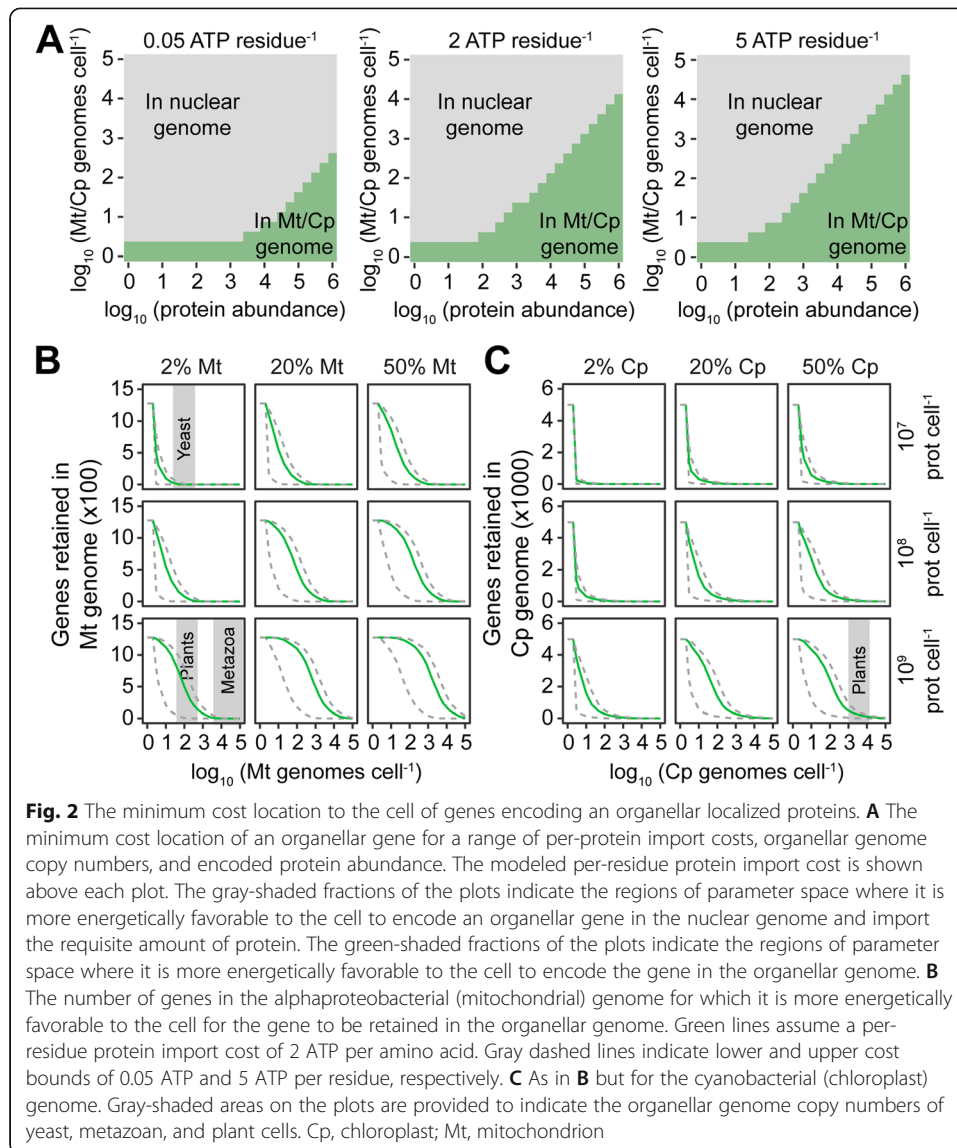
The cost to the cell to encode a gene in the organellar genome is higher than in the nuclear genome

Eukaryotic cells possessing chloroplasts and/or mitochondria typically have a higher copy number of their organellar genomes than their nuclear genomes [68]. Accordingly, while a typical diploid cell will have two copies of every gene in the nuclear genome, the same cell will have hundreds to hundreds of thousands of copies of every organellar encoded gene [68]. This difference in per-cell genome copy number means that it costs the cell more DNA to encode a gene in the organellar genome than in the nuclear genome. To provide an illustration of this difference in cost, three model eukaryotes were selected with disparate genome sizes and organellar genome content which are representative of the diverse range of values that have been previously reported [68]. Here, the cost of encoding a gene in a nuclear or organellar genome was considered to be the ATP cost of the chromosome (organellar or nuclear) divided by the number of genes on that chromosome. This consideration was performed to account for the differences in organellar and nuclear genomes such as the presence of introns, structural elements (telomeres, centromeres, etc.), and regulatory elements. We also included the ATP cost of the requisite number of histone proteins contained in nucleosomes to compute the cost of encoding a gene in the nuclear genome. This revealed that the high per-cell organellar genome copy number meant that the ATP cost of encoding a gene in the organellar genome is on average one order of magnitude higher than the cost of encoding a gene in the nuclear genome (Fig. 1A). This difference in ATP cost is further enhanced if the cost of just the coding sequences (including nucleosomes but excluding introns and non-coding regions) is compared directly (Fig. 1B). This latter scenario is more similar to a recent endosymbiotic gene transfer that arrives in the nuclear genome without introns and acquires these over time [72]. As the three representative organisms shown here span the range of organellar genome copy numbers that have been observed in eukaryotes [68], it follows that the ATP cost to the cell of encoding a gene in the organellar genome is generally higher than the cost of encoding the same gene in the nuclear genome in eukaryotes. Consequently, for any organellar gene, the cell may be able to save resources by transferring that gene from the organellar genome to the nuclear genome or by acquiring a functionally equivalent gene through lateral gene transfer and deleting the organellar gene.



The energy saved by encoding a gene in the nuclear genome instead of the organellar genome is sufficient to offset the cost of organellar protein import

Although it is cheaper for the cell to encode a gene in the nuclear genome than the organellar genome, this direct cost comparison only considers the cost of DNA (and its associated proteins) and does not account for the additional cost that would be incurred should the product of a nuclear-encoded gene be required to function in the organelle. Such nuclear-encoded organelle-targeted proteins incur additional energetic costs to be translocated across the organellar membranes. Accordingly, to assess whether it is cheaper for the cell to encode an organelle-targeted protein in the nuclear or organellar genome, it is necessary to consider both the abundance of the encoded protein and the energetic cost of organellar protein import. Estimates for the energetic cost of mitochondrial or chloroplast protein import vary over two orders of magnitude from ~0.05 ATP per amino acid to 5 ATP per amino acid [73–75]. Thus, for the purposes of this study, the full range of estimates was considered and the range of conditions under which it is more energetically favorable to encode a gene in the organellar or nuclear genome was assessed. This analysis revealed that the higher the copy

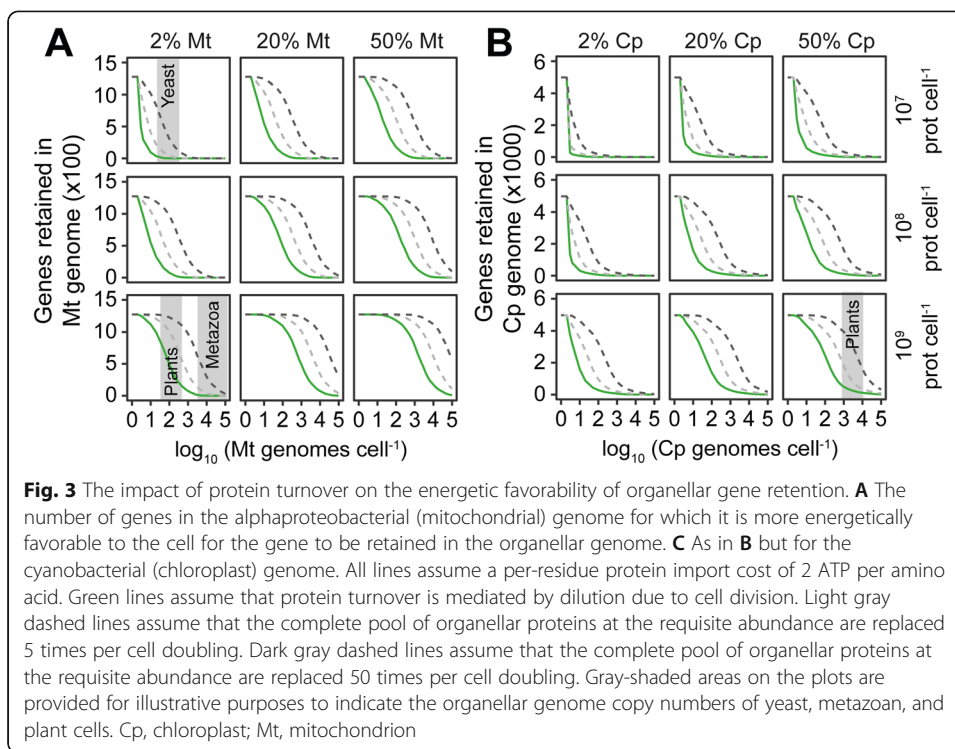


number of the organellar genome, the more energy that is saved by encoding the gene in the nuclear genome and thus the more protein that can be imported into the organelle while still reducing the overall energetic cost of the cell (Fig. 2A). As the per-cell gene copy number is the same for each gene encoded on the organellar genome, the possible energetic advantage to the cell arising from endosymbiotic gene transfer will vary between genes as a function of the required abundance of each encoded gene product. Furthermore, if the cell can function without the encoded gene product, then as organellar genome copy number increases the energetic incentive to discard the gene also increases. Thus, high organellar genome copy numbers provide an energetic incentive to either delete genes from the organellar genome or transfer them to the nuclear genome.

Given that the magnitude of the energetic advantage of endosymbiotic gene transfer is dependent on protein abundance, we sought to simulate the endosymbiotic genome reduction that would occur using realistic models of pre-mitochondrial and pre-

chloroplast organellar progenitors. Here, the complete genomes with measured protein abundances for an alphaproteobacterium (*Bartonella henselae*) and a cyanobacterium (*Microcystis aeruginosa*) were chosen as models for the mitochondrial and chloroplast progenitors, respectively. In addition, a range of host cell size (i.e., host cell protein content) was considered such that it encompassed the majority of diversity exhibited by extant eukaryotes [76] and would thus likely also encompass the size range of the host cell that originally engulfed the organellar progenitors. This range extended from a small unicellular yeast-like cell (10^7 proteins) to a large metazoan/plant cell (10^9 proteins). Each of these cell types was then considered to allocate a realistic range of total cellular protein to mitochondria/chloroplasts representative of values observed in extant eukaryotic cells (Additional file 2: Table S1). For each set of conditions in this comprehensive parameter space, the energy liberated or incurred by endosymbiotic gene transfer was calculated for each organellar gene given its measured protein abundance [77] and a realistic range of protein import costs (including the total biosynthetic cost of the protein import machinery, see the “Methods” section). This revealed that for a broad range of estimates of cell size, organellar genome copy number, and organellar fraction (i.e., the fraction of the total number of protein molecules in a cell that are contained within the organelle), it is energetically favorable to the cell to transfer the majority of organellar genes to the nuclear genome and re-import the proteins back to the organelle (Fig. 2B, C). Only the proteins with the highest abundance, and thus which incur the largest import cost, are energetically favorable to be retained in the organellar genomes. This phenomenon was also observed even if extreme costs for protein import ten times those that have been measured are considered (Fig. S1). Thus, it is more energy efficient for a eukaryotic cell to position the majority of genes that encode organellar targeted proteins in the nuclear genome.

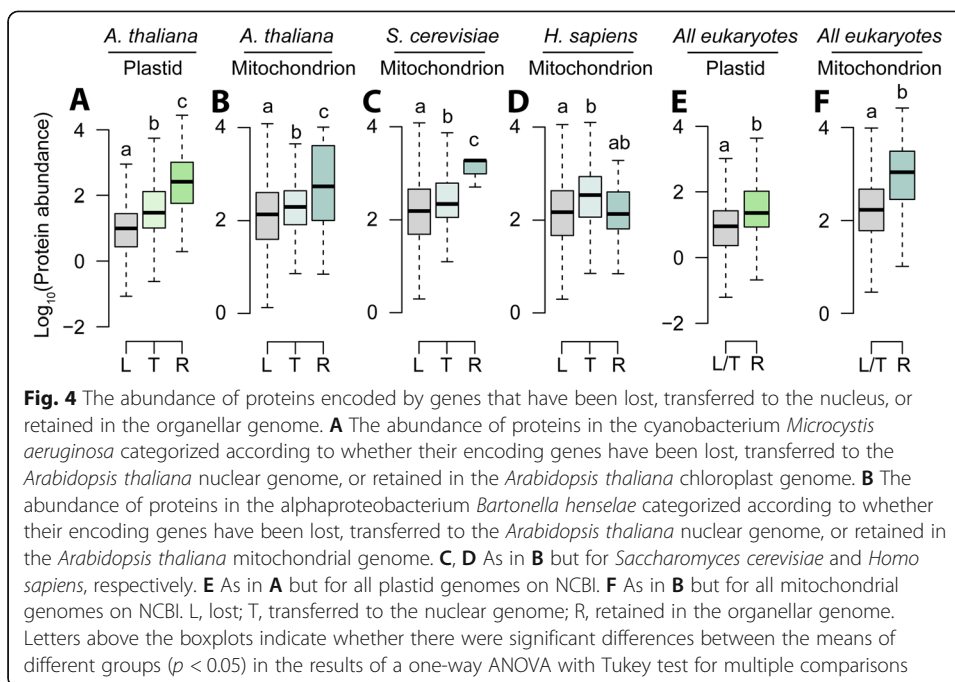
The above analysis assumed that the total pool of cellular protein was replaced with each cell doubling. This assumption is consistent with the observations that protein turnover in eukaryotes (as in bacteria) is primarily mediated by dilution due to cell division [78–80], i.e., the vast majority of proteins have half-lives that are longer in duration than the doubling time of the cell, and thus protein turnover occurs through replicative dilution. However, a small population of proteins is turned over more than once per cell division cycle [78–80], and in multicellular organisms, there can be populations of cells with a low or negligible rate of cell division resulting in a higher rate of protein turnover per cell division. Similarly, some of the archaeal relatives of the last eukaryotic common ancestor have slow cell doubling rates and thus may have higher rates of protein turnover relative to cell doubling. Thus, to determine the impact of enhanced rates of protein turnover relative to cell doubling, the analysis above was repeated while increasing the rate of protein turnover from once per cell division cycle (i.e., dividing cells) to 50 times per cell doubling (i.e., a long-lived or non-dividing cell). Increasing the rate of protein turnover increases the total amount of protein that must be imported into the organelle (akin to an increase in absolute abundance of that protein) and thus leads to an increase in the number of proteins for which it is energetically favorable to retain their corresponding genes in the organellar genomes (Fig. 3A, B). However, even if it is assumed that the total pool of each organellar protein is turned over 50 times per cell doubling, it is still more energetically favorable to transfer the majority of organellar genes to the



nuclear genome when the organellar genome copy number is high (Fig. 3A, B). Thus, in both dividing cells and in cells with higher rates of protein turnover relative to cell division, it is more energetically favorable to encode the majority of organellar targeted proteins in the nuclear genome.

Proteins encoded by organellar genes have higher estimated ancestral abundance than those that have been lost or transferred to the nuclear genome

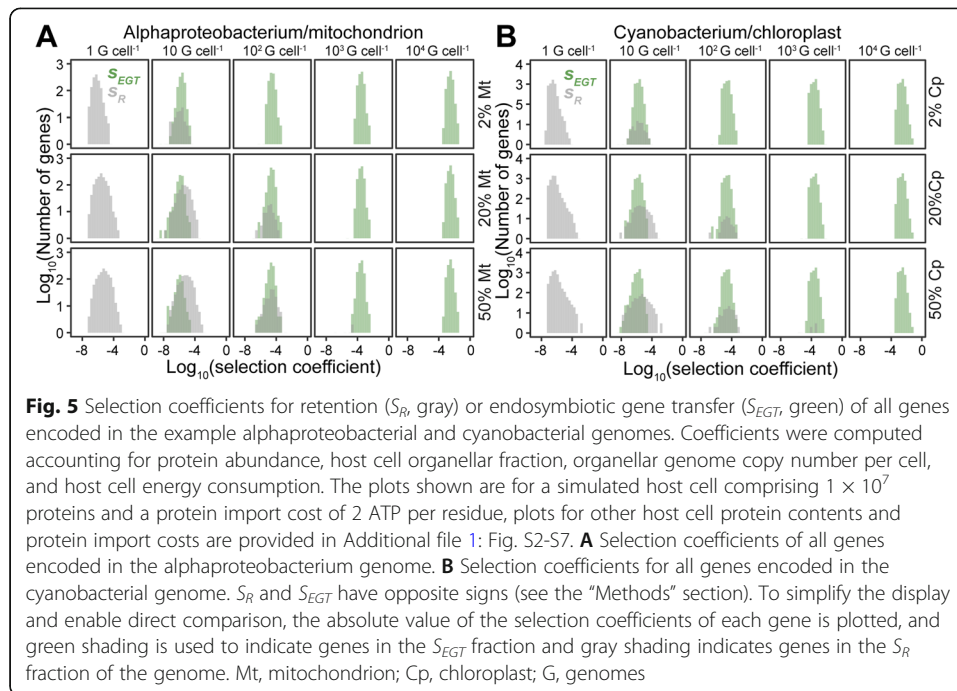
The analyses above predict that the proteins with the highest abundance, and thus those which incur the highest import costs, are those that are more likely to be retained in an organellar genome. While it is unknown how the abundance of proteins in organelles has changed throughout the evolution of the eukaryotes, it is possible to estimate what the profile of protein abundances may have looked like during the initial stages of this process by examining protein abundance in extant bacterial relatives of organelles [77]. Using these inferred ancestral protein abundance estimates, it is thus possible to ask whether those genes that are retained in the organellar genome are those that encode proteins with higher abundance than those that are lost or transferred to the nuclear genome. This revealed that the estimated abundance of the cohorts of proteins whose genes are retained in the chloroplast (Fig. 4A) and mitochondrial (Fig. 4B) genomes of *Arabidopsis thaliana* and the mitochondrial genome of *Saccharomyces cerevisiae* (Fig. 4C) is significantly higher than the estimated abundance of the cohorts of proteins that were either lost or transferred to the respective nuclear genomes. The estimated abundance of the cohort of proteins whose genes are retained in the mitochondrial genome of *Homo sapiens* was not significantly different from those that have been lost or transferred to the nuclear genome (Fig. 4D). To assess whether or not this



elevated protein abundance was a general phenomenon, the full set of complete plastid and mitochondrial genomes were downloaded from NCBI, and the sets of genes present or absent from these genomes were analyzed. Here, the corresponding nuclear genomes were not available, so it was not possible to separately assess the estimated abundance proteins encoded by lost or putatively transferred genes, and thus, they were analyzed together. This analysis revealed that the estimated abundance of proteins encoded by genes found in the extant plastid (Fig. 4E) or mitochondrial (Fig. 4F) genomes in eukaryotes was significantly higher than those that have been lost or transferred to the nuclear genome. Thus, across all eukaryotes, the inferred ancestral abundance of proteins encoded by genes retained in organellar genomes is higher than those encoded by genes that were either lost or transferred to the nuclear genome.

The energy saved by gene loss or endosymbiotic gene transfer is sufficient to produce a selectable advantage for the majority of genes

Although gene loss or endosymbiotic gene transfer can save energy, the question arises as to whether this energy saving would be sufficient to confer a selectable advantage for the cell. To estimate this, the energy liberated by endosymbiotic gene transfer of each gene encoded in the ancestral pre-organellar genomes was evaluated as a proportion of the total energy required to replicate the cell. As above, this analysis was conducted for a broad range of host cell size, organellar fraction, endosymbiont/organellar genome copy number, and protein import cost that is representative of a broad range of eukaryotic cells (Fig. 5A, B; Figs. S2–S7). This revealed that for even modest per-cell endosymbiont genome copy numbers (~ 100 copies per cell), the proportion of the total cell energy budget that could be saved for an individual gene transfer event (or equivalent functional lateral complementation) is sufficient that it would confer a selectable advantage. If the energetic advantage is considered to be a direct fitness advantage then



the selection coefficients for the transfer of the majority of individual endosymbiont genes are $\sim 1 \times 10^{-5}$ (Fig. 5; Figs. S2–S7). This is ~ 1000 times stronger than the selection coefficient acting against disfavored synonymous codons [81]. Moreover, for high per-cell endosymbiont genome copy numbers (~ 1000 genome copies per cell), these selection coefficients are proportionally larger ($\sim 1 \times 10^{-4}$), equivalent to approximately 1/10th the strength of the selection that caused the allele conferring lactose tolerance to rapidly sweep through human populations in ~ 500 generations [82]. In contrast, selection coefficients for retention of genes in the organellar genome generally only occur when organellar genome copy numbers are low, and/or when large proportions of cellular resources are invested in the organelle (Fig. 5A, B; Figs. S2–S7). Consistent with the analysis of protein turnover relative to cell doubling time (Fig. 3), these results are recovered even for cells with ten times the cell doubling time considered here (Figs. S8–S13). Thus, over a broad range of host cell sizes, organellar genome copy numbers, organellar fractions, and per-protein ATP import costs, protein turn-over rates, and cell doubling times endosymbiotic gene transfer of the majority of genes is sufficiently energetically advantageous that any such transfer events, if they occurred, would confer an energetic advantage to the cell and have the potential to rapidly reach fixation (Fig. S14). Thus, endosymbiotic gene transfer of the majority of organellar genes is advantageous to eukaryotic cells.

Discussion

The endosymbiosis of the bacterial progenitors of the mitochondrion and the chloroplast are landmark events in the evolution eukaryotes. Following these endosymbioses, there was a dramatic reduction in the gene content of the organellar genomes such that they now harbor fewer than 5% of the genes found in their free-living bacterial relatives. Some of these genes have been discarded, but many have been transferred to the

nuclear genome and their products (proteins) are imported back into the organelle where they function. The reason why these organelles have transferred their genes to the nucleus is a long-standing unanswered question in evolutionary biology. Here, we show, through extensive simulation of plausible parameter spaces for eukaryotic cells, that there are energy incentives for gene loss and for endosymbiotic gene transfer from organellar genomes. We show that these energy incentives are dependent on the abundance of the encoded gene product, with a trade-off between per-cell organellar genome copy number and protein abundance determining the magnitude and direction of the energy incentive. We further show that these energy incentives can be sufficient to produce a selectable advantage to the host cell for both endosymbiotic gene transfer and retention of genes in the organellar genomes. Thus, the economics of protein production and transport plays a role in determining whether genes are lost, retained, or transferred from organellar genomes.

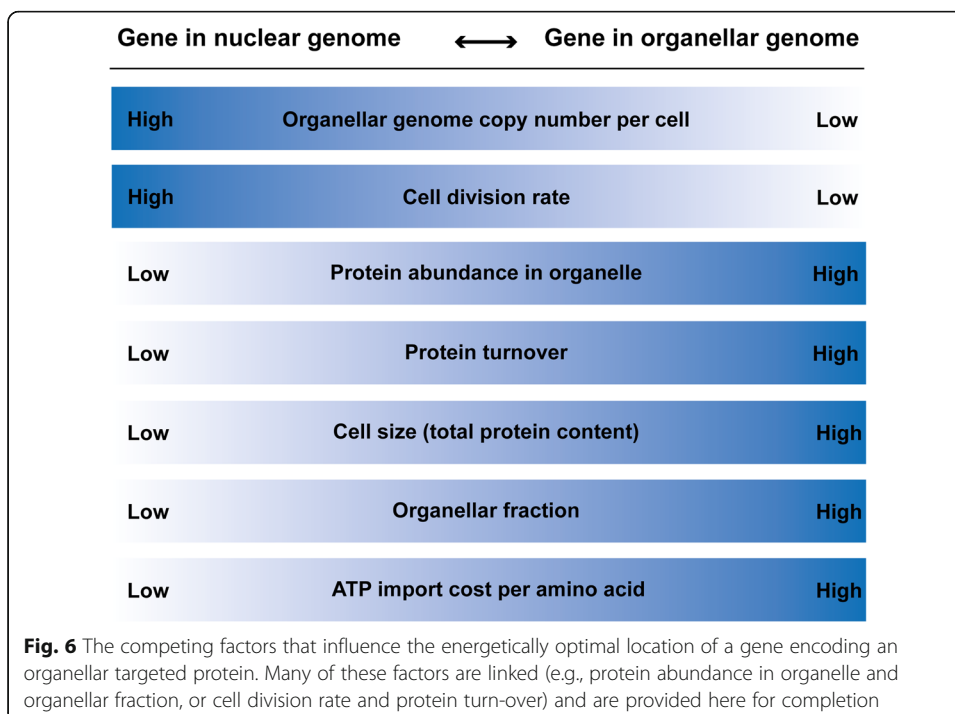
Although this study reveals that the energy efficiency of protein production can provide a driver for the location of an organellar gene, it is not proposed that it is the only factor that influences this process. Instead, a large cohort of factors including the requirement for organellar-mediated RNA editing, protein chaperones, protein folding, post-translational modifications, escaping mutation hazard, Muller's ratchet, enhanced nuclear control, the requirement for redox regulation of gene expression, and drift will act antagonistically or synergistically with energetic incentives described here to influence the set of genes that are retained in, lost, or transferred from, the organellar genomes. The study presented here simply reveals that energy efficiency is a previously overlooked factor that has likely played a role in shaping the evolution organellar/nuclear genomes. Moreover, the work presented here is in agreement, and is synergistic, with previous hypotheses that have suggested that the reason for retaining genes in organellar genomes is that there is a selectable advantage to do so. Specifically, the CoRR hypothesis [42, 63–65] posits that genes are retained in organellar genomes as it is advantageous for the cell to be able to control gene expression (and the gene products that are made) in immediate and direct response to the redox state of the organelle. These redox-regulated genes are also required in very high abundance within the organelle, and thus, the selection on energetic incentives acts in the same direction as selection for maintaining tight redox regulation. Stochastic models of populations of cells in which endosymbiotic gene transfer (or functionally equivalent lateral gene transfer) is occurring may provide insight into the synergy and conflict between this diverse set of factors, and their relative contribution to the evolution of organellar genomes.

It is noteworthy in these contexts that if the protein encoded by the endosymbiont gene can provide its function outside of the endosymbiont (e.g., by catalyzing a reaction that could occur equally well in the cytosol of the host as in the endosymbiont), then the energetic advantage of gene transfer to the nuclear genome is further enhanced, as the cost of protein import is not incurred. Similarly, although gene loss has been proposed to be mediated predominantly by mutation pressure and drift [20], the elevated per-cell endosymbiont genome copy number also provides a substantial energetic reward to the host cell for complete gene loss as neither the costs of encoding the gene or producing its product are incurred. Thus, high organellar genome copy number

provides an energetic incentive for the cell to delete endosymbiont genes or transfer them to the nuclear genome.

While the analysis presented here focussed on the energetic cost measured in ATP so that the cost of protein import and the cost of biosynthesis of DNA could be evaluated on a common basis, endosymbiotic gene transfer also results in changes in the elemental requirements of a cell. Specifically, as the monophosphate nucleotides that constitute DNA are composed of carbon (A = 10, C = 9, G = 10, T = 10), nitrogen (A = 5, C = 3, G = 5, T = 2), and phosphorous (A = 1, C = 1, G = 1, T = 1) atoms, endosymbiotic gene transfer can also result in substantial savings of these resources (Fig. S15). Thus, if organisms encounter carbon, nitrogen, or phosphorous limitation in their diet and environment, then the advantage of endosymbiotic gene transfer to the cell will be further enhanced.

The analysis presented here shows that a broad range of cell sizes and resource allocations that endosymbiotic gene transfer of the majority of organellar genes is energetically favorable and thus advantageous to the cell. However, it also showed that retention of genes in the organellar genomes is energetically favorable under conditions where the encoded organellar protein is required in very high abundance and/or the copy number of the organellar genome is low. Other interlinked competing factors that influence the energetically optimal location of a gene are shown in Fig. 6. Each of these factors interacts to influence the cost to the cell for encoding a gene in the nuclear or organellar genome. This is important, as while we do not know precisely what the cells that engulfed the progenitors of the mitochondrion or the chloroplast looked like (as only extant derivatives survive), it is safe to assume that cell size and investment in organelles has altered since these primary endosymbioses first occurred. Accordingly, the selective advantage (or disadvantage) of transfer of any given gene is transient and will have varied during the radiation of the eukaryotes as factors such as cell size and



organellar volume evolved and changed in disparate eukaryotic lineages. This coupled with the lack of an organellar protein export system (i.e., from the organelle to the host cytosol) and the presence (and acquisition) of introns in nuclear-encoded genes [83] means that it is more difficult for endosymbiotic gene transfer to operate in the reverse direction (i.e., from the nucleus to organelle). Similarly, eukaryotic cells can typically tolerate the loss of one or more chloroplasts [84] or mitochondria [85] from a cell without the concomitant death of the cell, the disruption of these organelles is thought to be a major route through which DNA from organelles enters the nucleus and can thus be incorporated into the nuclear genome. The converse process (i.e., the loss of the nucleus) is terminal to the cell and is thought to be a major reason why endosymbiotic gene transfer operates in one direction only. Collectively, these factors would create ratchet-like effect trapping genes in the nuclear genome even if subsequent changes in cell size and organellar fraction means that it became energetically advantageous to return the gene to the organelle later in the evolution. Thus, current organellar and nuclear gene contents predominantly reflect past pressures to delete organellar genes or transfer them to the nuclear genome.

Conclusion

Endosymbiotic gene loss and gene transfer are a recurring theme in the evolution of the eukaryotic tree of life. The discovery that endosymbiotic gene transfer (or equivalent functional lateral complementation) can provide an energetic advantage to the cell for loss, retention, or transfer of organellar genes to the nuclear genome uncovers a novel process that has helped shape the content and evolution of eukaryotic genomes.

Methods

Data sources

The *Arabidopsis thaliana* genome sequence and the corresponding set of representative gene models were downloaded from Phytozome V13 [86]. The human genome sequence and gene models from assembly version GRCh38.p13 (GCA_000001405.28), the *Bartonella henselae* genome sequence and gene models from assembly version ASM4670v1, and the *Microcystis aeruginosa* NIES-843 genome sequence and gene models from assembly version ASM1062v1 were each downloaded from Ensembl [87]. The *Saccharomyces cerevisiae* sequence and gene models from assembly version R64-2-1_20150113 were downloaded from the *Saccharomyces* Genome Database [88]. Protein abundance data for all species were obtained from PAXdb v4.1 [77].

Constants used to evaluate the per cell ATP costs of genes and chromosomes

The ATP biosynthesis cost of nucleotides and amino acids was obtained from [89] and [71] and are provided in Additional file 3: Table S2. The *Homo sapiens* mitochondrial genome copy number of 5000 was obtained from [68]. The *Saccharomyces cerevisiae* mitochondrial genome copy number of 200 was obtained from [90]. The *Arabidopsis thaliana* chloroplast genome copy number of 1500 was obtained from [91], and the *Arabidopsis thaliana* mitochondrial genome copy number of 100 was obtained from [68].

For genes in nuclear chromosomes, the cost of DNA was calculated to include the cost of nucleosomes with one histone octamer comprising two copies each of the histone proteins H2A, H2B, H3, and H4 every 180 bp (147 bp for the two turns of DNA around the histone octamer and 33 bp for the spacer) [71]. For organellar chromosomes, there are no histones/nucleosomes, and thus, the biosynthetic cost of genes in organellar chromosomes was calculated as the cost of the DNA divided by the number of genes on the chromosome (Additional file 4: Table S3). Although there are no histone protein equivalents in that organellar genomes, it should be noted that there are some nuclear-encoded proteins that are known to bind mitochondrial or chloroplast DNA. The costs associated with these proteins have not been included here as their function in packaging DNA is unknown and their density within the organellar genome is also unknown, and it is thus difficult to estimate their required abundance. However, the inclusion of the production and import costs of these proteins would further increase the cost of encoding a gene in the organellar genome and would accentuate the differences shown in this study.

The average gene length used for the simulation study in Fig. 2 was obtained by computing the average gene length across the two bacterial genomes used in this study, *Bartonella henselae* ASM4670v1 and *Microcystis aeruginosa* NIES-843.

Calculating protein import costs

Although the molecular mechanisms of mitochondrial and chloroplast protein import differ [92–94], they share many commonalities including the requirement for energy in the form of nucleoside triphosphate hydrolysis [95]. The energetic cost of mitochondrial or chloroplast protein import is difficult to measure directly, and accordingly, estimates vary over two orders of magnitude from ~0.05 ATP per amino acid to 5 ATP per amino acid [73–75]. Thus, for the purposes of this study, the full range of estimates was considered in all simulations when evaluating the import cost of organellar targeted proteins encoded by nuclear genes.

The cost of the biosynthesis of the protein import machinery (i.e., the TOC/TIC or TOM/TIM complexes, Additional file 5: Table S4) was also included in the per protein import costs calculated in this study. For *Arabidopsis thaliana*, if the total ATP biosynthesis cost of all TOC/TIC complex proteins in the cell (i.e., the full biosynthesis cost of all the amino acids of all the proteins at their measured abundance in the cell) is distributed equally among all of the proteins that are imported into the chloroplast, then it would add an additional 0.2 ATP per residue imported (Additional file 6: Table S5). Similarly, if the total ATP biosynthesis cost of all TOM/TIM proteins in the cell in *Homo sapiens*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana* is distributed equally among all of the proteins that are imported into the mitochondrion in those species, then it would add an additional 0.2 ATP, 0.7 ATP, and 0.2 ATP per residue imported, respectively (Additional file 6: Table S5). In all cases, protein abundance was calculated using measured protein abundance estimates for each species obtained from PAXdb 4.0 [77], assuming a total cell protein content of 1×10^9 proteins for a human cell, 1×10^7 proteins for a yeast cell, and 2.5×10^{10} proteins for an *Arabidopsis thaliana* cell. As we modeled ATP import costs from 0.05 ATP to 50 ATP per residue, the

cost of the import machinery was considered to be included within the bounds considered in this analysis.

Evaluating the proportion of the total proteome invested in organelles

To provide estimates of the fraction of cellular protein resources invested in organellar proteomes, the complete predicted proteomes and corresponding protein abundances were quantified. Organellar targeting was predicted using TargetP-2.0 [96], and protein abundance estimates were obtained from PAXdb 4.0 [77]. The proportion of cellular resources are provided in Additional file 2: Table S1 and were used to provide the indicative regions or parameter space occupied by metazoa, yeast, and plants shown in Fig. 2B, C. Specifically, ~5% of total cellular protein is contained within mitochondria in *H. sapiens*, *S. cerevisiae*, and *A. thaliana*, and ~50% of total cellular protein is contained within chloroplasts in *A. thaliana*.

Calculating the free energy of endosymbiotic gene transfer

The free energy of endosymbiotic gene transfer (ΔE_{EGT}) is here defined as the difference in energy cost to the cell to encode a given gene in the organellar genome and the cost to encode the same gene in the nuclear genome and import the requisite amount of gene product into the organelle. ΔE_{EGT} is evaluated as the difference in ATP biosynthesis cost required to encode a gene (ΔD) in the endosymbiont genome (D_{end}) and the nuclear genome (D_{nuc}) minus the difference in ATP biosynthesis cost required to produce the protein (ΔP) in the organelle (P_{end}) vs in the cytosol (P_{cyt}) and ATP cost to import the protein into the organelle (P_{import}). Such that:

$$\Delta E_{EGT} = \Delta D - \Delta P \quad (1)$$

where

$$\Delta D = D_{end} - D_{nuc} \quad (2)$$

and

$$\Delta P = P_{end} - P_{cyt} - P_{import} \quad (3)$$

Thus, ΔE_{EGT} can be positive or negative depending on the cost associated with each parameter. The energetic cost of producing a protein in the endosymbiont and in the cytosol is assumed to be equal, and thus:

$$\Delta P = P_{import} \quad (4)$$

P_{import} is evaluated as the product of the length of the amino acid sequence (L_{prot}), the ATP cost of importing a single residue from the contiguous polypeptide chain of that protein (C_{import}), and the number of copies of that protein contained within the cell that must be imported (N_p) such that:

$$\Delta P = P_{import} = L_{prot} C_{import} N_p \quad (5)$$

Measured estimates of C_{import} range from ~0.05 ATP per amino acid to 5 ATP per amino acid [73–75]. For the purposes of this study, we used these measured ranges and also modeled a C_{import} up to 10 times higher than any measured estimate, i.e., from 0.05 ATP to 50 ATP.

Both D_{end} and D_{nuc} are evaluated as the product of the ATP biosynthesis cost of the double-stranded DNA (A_{DNA}) that comprises the gene under consideration and the copy number (C) of the genome in the cell such that:

$$D_{\text{end}} = A_{\text{DNA}} C_{\text{end}} \quad (6)$$

And

$$D_{\text{nuc}} = A_{\text{DNA}} C_{\text{nuc}} \quad (7)$$

Such that:

$$\Delta D = A_{\text{DNA}} (C_{\text{end}} - C_{\text{nuc}}) \quad (8)$$

where C_{end} and C_{nuc} are the per-cell copy number of the endosymbiont and nuclear genomes, respectively, and the ATP biosynthesis cost for the complete biosynthesis of an A:T base pair and a G:C base pair is 40.55 ATP and 40.14 ATP, respectively [89]. Thus:

$$\Delta E_{\text{EGT}} = A_{\text{DNA}} (C_{\text{end}} - C_{\text{nuc}}) - L_{\text{prot}} C_{\text{import}} N_p \quad (9)$$

where positive values of ΔE_{EGT} correspond to genes for which it is more energetically favorable to be encoded in the nuclear genome, and negative values correspond to genes for which it is more energetically favorable to be encoded in the endosymbiont genome. Other studies have used slightly higher estimates (~ 50 ATP per nucleotide) for the biosynthesis cost of nucleotides [71, 97]. However, as this value is always used in the product with the difference in per-cell copy number of the endosymbiotic and nuclear genomes [8, 9], this would have a marginal effect on the results of the models. This is because the difference in copy number ranges over 5 orders of magnitude while the difference in the estimates of nucleotide biosynthesis cost varies by 20%.

Simulating endosymbiotic gene transfer of mitochondrial and chloroplast genes

The complete genomes with measured protein abundances for an alphaproteobacterium (*Bartonella henselae*) and a cyanobacterium (*Microcystis aeruginosa*) were selected to serve as models for an ancestral mitochondrion and cyanobacterium, respectively. To account for uncertainty in the size and complexity of the ancestral pre-mitochondrial and pre-chloroplast host cells, a range of potential ancestral cells was considered to be engulfed by a range of different host cells with protein contents representative of the diversity of extant eukaryotic cells [76]. Specifically, the size of the host cell ranged from a small unicellular yeast-like cell (10^7 proteins) to a medium-sized unicellular algal-like cell (10^8 proteins) to a typical metazoan/plant cell (10^9 proteins). Each of these host cell types was then considered to allocate a realistic range of total cellular protein to mitochondria/chloroplasts typical of eukaryotic cells (i.e., $\sim 2\%$ for yeast [98], $\sim 20\%$ for metazoan cells [99], and $\sim 50\%$ of the non-vacuolar volume of plant cells [100]). It is not important whether the organellar fraction of the cell is composed of a single large organelle or multiple smaller organelles as all costs, abundances, and copy numbers are evaluated at a per-cell level. For each simulated cell, ΔE_{EGT} was evaluated for each gene in the endosymbiont genome using real protein abundance data [77] for a realistic range of endosymbiont genome copy numbers using Eq. 9. In all cases, the host cell was assumed to be diploid. The simulations were repeated for three different

per-residue protein import costs (0.05 ATP, 2 ATP, and 5 ATP per residue). The number of genes where ΔE_{EGT} was positive was recorded as these genes comprise the cohort that is energetically favorable to be encoded in the nuclear genome.

Estimating the strength of selection acting on endosymbiotic gene transfer

To model the proportion of energy that would be saved by an individual endosymbiotic gene transfer event, a number of assumptions were made. It was assumed that the ancestral host cell had a cell size that is within the range of extant eukaryotes (i.e., between 1×10^7 proteins per cell and 1×10^9 proteins per cell). It was assumed that the endosymbiont occupied a fraction of the total cell proteome that is within the range exhibited by most eukaryotes today (2 to 50% of total cellular protein is located within the endosymbiont under consideration). It was assumed that endosymbiont genome copy number ranged between 1 copy per cell (as it most likely started out with a single copy) and 10,000 copies per cell.

We assumed an ancestral host cell with a 24-h doubling time such that all genomes and proteins are produced in the required abundance every 24-h period. As previously defined [71], the energy required for cell growth was modeled as:

$$C_r = 26.92V^{0.97} \quad (10)$$

In addition, all cells, irrespective of whether they are bacterial or eukaryotic, consume ATP (C_m) in proportion to their cell volume (V) at approximately the rate of:

$$C_m = 0.39V^{0.88} \quad (11)$$

where C_m is in units of 10^9 molecules of ATP $\text{cell}^{-1} \text{h}^{-1}$, and V is in units of μm^3 [71]. Thus, the total energy (E_R) needed to replicate a cell was considered to be:

$$E_R = C_r + 24 C_m \quad (12)$$

The proportional energetic advantage or disadvantage ($E_{A/D}$) to the host cell from the endosymbiotic gene transfer of a given gene is evaluated as the free energy of endosymbiotic gene transfer divided by the total amount of energy consumed by the cell during its 24-h life cycle.

$$E_{A/D} = \frac{\Delta E_{EGT}}{E_R} \quad (13)$$

Given that $E_{A/D}$ describes the proportional energetic advantage or disadvantage a cell has from a given endosymbiotic gene transfer event $E_{A/D}$ can be used directly as selection coefficient (s) to evaluate the strength of selection acting on the endosymbiotic gene transfer of a given gene, such that:

$$s = E_{A/D} \quad (14)$$

As ΔE_{EGT} can be positive or negative as described above, s is therefore also positive or negative depending on the endosymbiont genome copy number, endosymbiont fraction, host cell protein content, the abundance of the protein that must be imported, and the ATP cost of protein import. When s is less than 0, the absolute value of s is taken to be the selection coefficient for retention of a gene in the endosymbiont

genome (S_R); when s is greater than 0, the value of s is taken to be the selection coefficient for endosymbiotic gene transfer to the nucleus (S_{EGT}).

Estimating time to fixation

Fixation times for endosymbiotic gene transfer events for a range of observed selection coefficients from 1×10^{-5} to 1×10^{-2} were estimated using a Wright-Fisher model with selection and drift [101, 102] implemented in a simple evolutionary dynamics simulation [103]. The effective population size for these simulations was set as 1×10^7 , as is representative of unicellular eukaryotes [104], and multicellularity in eukaryotes is not thought to have evolved until after the endosymbiosis of either the mitochondrion or the chloroplast.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02567-w>.

Additional file 1. This file contains the 15 supplemental figures and their associated legends.

Additional file 2: Table S1. Proportion of total protein allocated to organelles.

Additional file 3: Table S2. The ATP biosynthesis costs of nucleotides and amino acids used in this study.

Additional file 4: Table S3. The cost of encoding genes in organellar vs nuclear chromosomes.

Additional file 5: Table S4. The protein components of the organellar protein import complexes.

Additional file 6: Table S5. The additional per-residue costs of including the protein import machinery.

Additional file 7. Review history

Acknowledgements

SK would like to thank Thomas A. Richards, David M. Emms, and John M. Archibald for their comments on the manuscript.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 7.

Author's contributions

SK conceived the study, conducted the experiments, and wrote the manuscript. The author read and approved the final manuscript.

Author's information

Twitter handle: @Steve__Kelly (Steven Kelly)

Funding

This work was funded by the Royal Society and the European Union's Horizon 2020 research and innovation program under grant agreement number 637765.

Availability of data and materials

The *Arabidopsis thaliana* genome sequence and the corresponding set of representative gene models were downloaded from Phytozome V13 [86]. The human genome sequence and gene models from assembly version GRCh38.p13 (GCA_000001405.28), the *Bartonella henselae* genome sequence and gene models from assembly version ASM4670v1, and the *Microcystis aeruginosa* NIES-843 genome sequence and gene models from assembly version ASM1062v1 were each downloaded from Ensembl [87]. The *Saccharomyces cerevisiae* sequence and gene models from assembly version R64-2-1_20150113 were downloaded from the *Saccharomyces* Genome Database [88]. Protein abundance data for all species were obtained from PAXdb v4.1 [77].

Declarations

Ethics approval and consent to participate

NA.

Competing interests

The author declares that there are no competing interests.

Received: 9 September 2021 Accepted: 6 December 2021

Published online: 20 December 2021

References

- Martin WF, Garg S, Zimorski V. Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci*. 2015; 370(1678):20140330. <https://doi.org/10.1098/rstb.2014.0330>.
- Archibald JM. Endosymbiosis and eukaryotic cell evolution. *Current Biology*. 2015;25(19):R911–R21. <https://doi.org/10.1016/j.cub.2015.07.055>.
- Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. Mitochondrial origins. *Proceedings of the National Academy of Sciences*. 1985;82(13):4443–7. <https://doi.org/10.1073/pnas.82.13.4443>.
- Roger AJ, Muñoz-Gómez SA, Kamikawa R. The origin and diversification of mitochondria. *Curr Biol*. 2017;27(21):R1177–r92. <https://doi.org/10.1016/j.cub.2017.09.015>.
- Martin W, Müller M. The hydrogen hypothesis for the first eukaryote. *Nature*. 1998;392(6671):37–41. <https://doi.org/10.1038/32096>.
- Martin W, Kowallik K. Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. *European Journal of Phycology*. 1999;34(3):287–95. <https://doi.org/10.1080/09670269910001736342>.
- Archibald JM. Genomic perspectives on the birth and spread of plastids. *Proceedings of the National Academy of Sciences*. 2015;112(33):10147–53. <https://doi.org/10.1073/pnas.1421374112>.
- Lane N, Martin WF. Eukaryotes really are special, and mitochondria are why. *Proceed Natl Acad Sci*. 2015;112(35):E4823–E.
- Lane N, Martin W. The energetics of genome complexity. *Nature*. 2010;467(7318):929–34. <https://doi.org/10.1038/nature09486>.
- Lane N. Bioenergetic constraints on the evolution of complex life. *Cold Spring Harb Perspect Biol*. 2014;6(5):a015982. <https://doi.org/10.1101/cshperspect.a015982>.
- Booth A, Doolittle WF. Eukaryogenesis, how special really? *Proceedings of the National Academy of Sciences*. 2015; 112(33):10278–85. <https://doi.org/10.1073/pnas.1421376112>.
- Booth A, Doolittle WF. Reply to Lane and Martin: Being and becoming eukaryotes. *Proceed Natl Acad Sci*. 2015;112(35):E4824–E.
- Lynch M, Marinov GK. Membranes, energetics, and evolution across the prokaryote-eukaryote divide. *eLife*. 2017;6: e20437. <https://doi.org/10.7554/eLife.20437>.
- Lynch M, Marinov GK. Response to Martin and colleagues: Mitochondria do not boost the bioenergetic capacity of eukaryotic cells. *Biology Direct*. 2018;13(1):26. <https://doi.org/10.1186/s13062-018-0228-3>.
- Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*. 2018;115(25):6506–11. <https://doi.org/10.1073/pnas.1711842115>.
- Gray MW, Burger G, Lang BF. Mitochondrial evolution. *Science*. 1999;283(5407):1476–81. <https://doi.org/10.1126/science.283.5407.1476>.
- Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet*. 2004;5(2):123–35. <https://doi.org/10.1038/nrg1271>.
- Green BR. Chloroplast genomes of photosynthetic eukaryotes. *Plant J*. 2011;66(1):34–44. <https://doi.org/10.1111/j.1365-3113.2011.04541.x>.
- McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*. 2012;10(1):13–26. <https://doi.org/10.1038/nrmicro2670>.
- Lynch M, Koskella B, Schaack S. Mutation pressure and the evolution of organelle genomic architecture. *Science*. 2006; 311(5768):1727–30. <https://doi.org/10.1126/science.1118884>.
- Smith DR. The mutational hazard hypothesis of organelle genome evolution: 10 years on. *Mol Ecol*. 2016;25(16):3769–75. <https://doi.org/10.1111/mec.13742>.
- Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences*. 2015;112(33):10177–84. <https://doi.org/10.1073/pnas.1422049112>.
- Brown JR. Ancient horizontal gene transfer. *Nature Reviews Genetics*. 2003;4(2):121–32. <https://doi.org/10.1038/nrg1000>.
- Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, et al. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol Evol*. 2013;5(1): 31–44. <https://doi.org/10.1093/gbe/evs117>.
- Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, et al. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Molecular Biology and Evolution*. 2008;25(4):748–61. <https://doi.org/10.1093/molbev/msn022>.
- Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences*. 2002;99(19):12246–51. <https://doi.org/10.1073/pnas.182432999>.
- Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol*. 2012;4(4):466–85. <https://doi.org/10.1093/gbe/evs018>.
- Calvo SE, Mootha VK. The mitochondrial proteome and human disease. *Annu Rev Genomics Hum Genet*. 2010;11(1):25–44. <https://doi.org/10.1146/annurev-genom-082509-141720>.
- Ferro M, Brugièrè S, Salvi D, Seigneurin-Berny D, Court M, Moyet L, et al. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol Cell Proteomics*. 2010;9(6):1063–84. <https://doi.org/10.1074/mcp.M900325-MCP200>.
- Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, et al. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*. 2013;153(7):1567–78. <https://doi.org/10.1016/j.cell.2013.05.040>.

31. Nakayama T, Ishida K. Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. *Curr Biol*. 2009;19(7):R284–5. <https://doi.org/10.1016/j.cub.2009.02.043>.
32. Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, Boo SM, et al. Differential gene retention in plastids of common recent origin. *Mol Biol Evol*. 2010;27(7):1530–7. <https://doi.org/10.1093/molbev/msq032>.
33. Nowack ECM, Vogel H, Groth M, Grossman AR, Melkonian M, Glöckner G. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Mol Biol Evol*. 2010;28(1):407–22. <https://doi.org/10.1093/molbev/msq209>.
34. Singer A, Poschmann G, Mühlich C, Valadez-Cano C, Hänisch S, Hüren V, et al. Massive protein import into the early-evolutionary-stage photosynthetic organelle of the amoeba *Paulinella chromatophora*. *Curr Biol*. 2017;27(18):2763–73.e5.
35. Nowack ECM, Weber APM. Genomics-informed insights into endosymbiotic organelle evolution in photosynthetic eukaryotes. *Ann Rev Plant Biol*. 2018;69(1):51–84. <https://doi.org/10.1146/annurev-arplant-042817-040209>.
36. Nowack EC, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. *Proc Natl Acad Sci U S A*. 2016;113(43):12214–9. <https://doi.org/10.1073/pnas.1608016113>.
37. Daley DO, Whelan J. Why genes persist in organelle genomes. *Genom Biol*. 2005;6(5):110. <https://doi.org/10.1186/gb-2005-6-5-110>.
38. Herrmann R. Eukaryotism, towards a new interpretation. *Eukaryotism and symbiosis*: Springer; 1997. p. 73–118, DOI: https://doi.org/10.1007/978-3-642-60885-8_7.
39. Martin W, Herrmann RG. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol*. 1998;118(1):9–17. <https://doi.org/10.1104/pp.118.1.9>.
40. Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr Biol*. 2006;16(23):2320–5. <https://doi.org/10.1016/j.cub.2006.09.063>.
41. Speijer D, Hammond M, Lukeš J. Comparing early eukaryotic integration of mitochondria and chloroplasts in the light of internal ROS challenges: timing is of the essence. *mBio*. 2020;11(3):e00955–20. <https://doi.org/10.1128/mBio.00955-20>.
42. Allen JF, Raven JA. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. *J Mol Evol*. 1996;42(5):482–92. <https://doi.org/10.1007/BF02352278>.
43. Muller HJ. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*. 1964;1(1):2–9. [https://doi.org/10.1016/0027-5107\(64\)90047-8](https://doi.org/10.1016/0027-5107(64)90047-8).
44. Lynch M. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol*. 1996;13(1):209–20. <https://doi.org/10.1093/oxfordjournals.molbev.a025557>.
45. Neiman M, Taylor DR. The causes of mutation accumulation in mitochondrial genomes. *Proceedings of the Royal Society B: Biological Sciences*. 2009;276(1660):1201–9. <https://doi.org/10.1098/rspb.2008.1758>.
46. Doolittle WF. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet*. 1998;14(8):307–11. [https://doi.org/10.1016/S0168-9525\(98\)01494-2](https://doi.org/10.1016/S0168-9525(98)01494-2).
47. Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology*. 2005;138(3):1723–33. <https://doi.org/10.1104/pp.105.060327>.
48. Hazkani-Covo E, Martin WF. Quantifying the number of independent organelle DNA insertions in genome evolution and human health. *Genome Biology and Evolution*. 2017;9(5):1190–203. <https://doi.org/10.1093/gbe/evx078>.
49. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics*. 2010;6(2):e1000834. <https://doi.org/10.1371/journal.pgen.1000834>.
50. Martin W. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proceedings of the National Academy of Sciences*. 2003;100(15):8612–4. <https://doi.org/10.1073/pnas.1633606100>.
51. Reyes-Prieto A. The basic genetic toolkit to move in with your photosynthetic partner. *Frontiers in Ecology and Evolution*. 2015;3(100).
52. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceed Natl Acad Sci*. 1987;84(24):9054–8. <https://doi.org/10.1073/pnas.84.24.9054>.
53. Smith DR. Mutation rates in plastid genomes: they are lower than you might think. *Genome Biol Evol*. 2015;7(5):1227–34. <https://doi.org/10.1093/gbe/evz069>.
54. Lynch M, Lynch PSTSM, Walsh B. *The origins of genome architecture*: Oxford University Press, Incorporated; 2007.
55. Grisdale CJ, Smith DR, Archibald JM. Relative mutation rates in nucleomorph-bearing algae. *Genom Biol Evol*. 2019;11(4):1045–53. <https://doi.org/10.1093/gbe/evz056>.
56. Drouin G, Daoud H, Xia J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol Phylogenet Evol*. 2008;49(3):827–31. <https://doi.org/10.1016/j.ympev.2008.09.009>.
57. Lynch M. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. *Mol Biol Evol*. 1997;14(9):914–25. <https://doi.org/10.1093/oxfordjournals.molbev.a025834>.
58. Khakhlova O, Bock R. Elimination of deleterious mutations in plastid genomes by gene conversion. *The Plant Journal*. 2006;46(1):85–94. <https://doi.org/10.1111/j.1365-3113X.2006.02673.x>.
59. Gallaher SD, Craig RJ, Ganesan I, Purvine SO, McCorkle SR, Grimwood J, et al. Widespread polycistronic gene expression in green algae. *Proceed Natl Acad Sci*. 2021;118(7):e2017714118. <https://doi.org/10.1073/pnas.2017714118>.
60. Guiliano DB, Blaxter ML. Operon conservation and the evolution of trans-splicing in the phylum Nematoda. *PLoS Genet*. 2006;2(11):e198. <https://doi.org/10.1371/journal.pgen.0020198>.
61. Michaeli S. Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. *Future Microbiol*. 2011;6(4):459–74. <https://doi.org/10.2217/fmb.11.20>.
62. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One*. 2015;10(7):e0132628. <https://doi.org/10.1371/journal.pone.0132628>.
63. Allen JF. Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. *Journal of Theoretical Biology*. 1993;165(4):609–31. <https://doi.org/10.1006/jtbi.1993.1210>.
64. Allen JF. Why chloroplasts and mitochondria retain their own genomes and genetic systems: collocation for redox regulation of gene expression. *Proc Natl Acad Sci U S A*. 2015;112(33):10231–8. <https://doi.org/10.1073/pnas.1500012112>.

65. Allen JF, Martin WF. Why have organelles retained genomes? *Cell Systems*. 2016;2(2):70–2. <https://doi.org/10.1016/j.cels.2016.02.007>.
66. Johnston IG, Williams BP. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Syst*. 2016;2(2):101–11. <https://doi.org/10.1016/j.cels.2016.01.013>.
67. Giannakis K, Arrowsmith SJ, Richards L, Gasparini S, Chustecki JM, Røyrvik EC, et al. Universal features shaping organelle gene retention. *bioRxiv*. 2021:2021.10.27.465964.
68. Cole LW. The evolution of per-cell organelle number. *Front Cell Dev Biol*. 2016;4:85. <https://doi.org/10.3389/fcell.2016.00085>.
69. Bendich AJ. Why do chloroplasts and mitochondria contain so many copies of their genome? *Bioessays*. 1987;6(6):279–82. <https://doi.org/10.1002/bies.950060608>.
70. Shokolenko I, Venediktova N, Bochkareva A, Wilson GL, Alexeyev MF. Oxidative stress induces degradation of mitochondrial DNA. *Nucleic Acids Res*. 2009;37(8):2539–48. <https://doi.org/10.1093/nar/gkp100>.
71. Lynch M, Marinov GK. The bioenergetic costs of a gene. *Proceed Natl Acad Sci*. 2015;112(51):15690–5. <https://doi.org/10.1073/pnas.1514974112>.
72. Ahmadinejad N, Dagan T, Gruenheit N, Martin W, Gabaldón T. Evolution of spliceosomal introns following endosymbiotic gene transfer. *BMC Evol Biol*. 2010;10(1):57. <https://doi.org/10.1186/1471-2148-10-57>.
73. Backes S, Herrmann JM. Protein translocation into the intermembrane space and matrix of mitochondria: mechanisms and driving forces. *Frontiers in Molecular Biosciences*. 2017;4(83).
74. Shi L-X, Theg SM. Energetic cost of protein import across the envelope membranes of chloroplasts. *Proceedings of the National Academy of Sciences*. 2013;110(3):930–5. <https://doi.org/10.1073/pnas.1115886110>.
75. Mokranjac D, Neupert W. Energetics of protein translocation into mitochondria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. 2008;1777(7):758–62. <https://doi.org/10.1016/j.bbabi.2008.04.009>.
76. Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*. 2013;35(12):1050–5. <https://doi.org/10.1002/bies.201300066>.
77. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 2015;15(18):3163–8. <https://doi.org/10.1002/pmhc.201400441>.
78. Boisvert FM, Ahmad Y, Gierliński M, Charrière F, Lamont D, Scott M, et al. A quantitative spatial proteomics analysis of proteome turnover in human cells. *Mol Cell Proteomics*. 2012;11(3):M111.011429. <https://doi.org/10.1074/mcp.M111.011429>.
79. Gawron D, Ndah E, Gevaert K, Van Damme P. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol Syst Biol*. 2016;12(2):858. <https://doi.org/10.15252/msb.20156662>.
80. Martin-Perez M, Villén J. Determinants and regulation of protein turnover in yeast. *Cell Syst*. 2017;5(3):283–94.e5.
81. Hartl DL, Moriyama EN, Sawyer SA. Selection intensity for codon bias. *Genetics*. 1994;138(1):227–34. <https://doi.org/10.1093/genetics/138.1.227>.
82. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 2004;74(6):1111–20. <https://doi.org/10.1086/421051>.
83. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. *Biol Direct*. 2012;7(1):11. <https://doi.org/10.1186/1745-6150-7-11>.
84. Zhuang X, Jiang L. Chloroplast degradation: multiple routes into the vacuole. *Frontiers in Plant Science*. 2019;10(359).
85. Ding WX, Yin XM. Mitophagy: mechanisms, pathophysiological roles, and analysis. *Biol Chem*. 2012;393(7):547–64. <https://doi.org/10.1515/hsz-2012-0119>.
86. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2012;40(Database issue):D1178–86. <https://doi.org/10.1093/nar/gkr944>.
87. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. *Nucleic Acids Res*. 2020;48(D1):D682–d8. <https://doi.org/10.1093/nar/gkz966>.
88. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*. 2012;40(Database issue):D700–5. <https://doi.org/10.1093/nar/gkr1029>.
89. Chen W-H, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nature Communications*. 2016;7(1):11334. <https://doi.org/10.1038/ncomms11334>.
90. Miyakawa I. Organization and dynamics of yeast mitochondrial nucleoids. *Proc Jpn Acad Ser B Phys Biol Sci*. 2017;93(5):339–59. <https://doi.org/10.2183/pjab.93.021>.
91. Zoschke R, Liere K, Börner T. From seedling to mature plant: Arabidopsis plastidial genome copy number, RNA accumulation and transcription are differentially regulated during leaf development. *The Plant Journal*. 2007;50(4):710–22. <https://doi.org/10.1111/j.1365-313X.2007.03084.x>.
92. Wiedemann N, Pfanner N. Mitochondrial machineries for protein import and assembly. *Annual Review of Biochemistry*. 2017;86(1):685–714. <https://doi.org/10.1146/annurev-biochem-060815-014352>.
93. Jarvis P. Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytologist*. 2008;179(2):257–85. <https://doi.org/10.1111/j.1469-8137.2008.02452.x>.
94. Soll J, Schleiff E. Protein import into chloroplasts. *Nat Rev Mol Cell Biol*. 2004;5(3):198–208. <https://doi.org/10.1038/nrm1333>.
95. Schatz G, Dobberstein B. Common principles of protein translocation across membranes. *Science*. 1996;271(5255):1519–26. <https://doi.org/10.1126/science.271.5255.1519>.
96. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance*. 2019;2(5).
97. Wagner A. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution*. 2005;22(6):1365–74. <https://doi.org/10.1093/molbev/msi126>.
98. Uchida M, Sun Y, McDermott G, Knoechel C, Le Gros MA, Parkinson D, et al. Quantitative analysis of yeast internal architecture using soft X-ray tomography. *Yeast*. 2011;28(3):227–36. <https://doi.org/10.1002/yea.1834>.
99. David H. Quantitative ultrastructural data of animal and human cells: Gustav Fischer; 1977.

100. Winter H, Robinson DG, Heldt HW. Subcellular volumes and metabolite concentrations in spinach leaves. *Planta*. 1994; 193(4):530–5. <https://doi.org/10.1007/BF02411558>.
101. Wright S. Evolution in Mendelian populations. *Genetics*. 1931;16(2):97–159. <https://doi.org/10.1093/genetics/16.2.97>.
102. Fisher RAS. The genetical theory of natural selection. Oxford: Clarendon Press; 1930. <https://doi.org/10.5962/bhl.title.27468>.
103. Niklaus M, Kelly S. The molecular evolution of C4 photosynthesis: opportunities for understanding and improving the world's most productive plants. *J Exper Botany*. 2018;70(3):795–804. <https://doi.org/10.1093/jxb/ery416>.
104. Lynch M, Conery JS. The origins of genome complexity. *Science*. 2003;302(5649):1401–4. <https://doi.org/10.1126/science.1089370>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

