Title

The economics of endosymbiotic gene transfer and the evolution of organellar genomes

One sentence summary

The high copy number of organellar genomes renders endosymbiotic gene transfer energetically favourable for the vast majority of organellar genes.

Authors

Steven Kelly

Affiliations

1) Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB

ORCID ID

SK: 0000-0001-8583-5362

Corresponding author

Name: Steven Kelly

Email: steven.kelly@plants.ox.ac.uk

Telephone: +44 (0)1865 275123

Abstract

The endosymbiosis of the bacterial progenitors of 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 genes found in living relatives of their ancestors. While some of the 95% of missing organellar genes have been discarded, many have been transferred to the host nuclear genome through a process known as endosymbiotic gene transfer. Here we demonstrate that the energy liberated or consumed by a cell as a result of endosymbiotic gene transfer can be sufficient to provide a selectable advantage for retention or nuclear-transfer of organellar genes in eukaryotic cells. We further demonstrate that for realistic estimates of protein abundances, organellar protein import costs, host cell sizes, and cellular investment in organelles that it is energetically favourable to transfer the majority of organellar genes to the nuclear genome. Moreover, we show that the selective advantage of such transfers is sufficiently large to enable such events to rapidly reach fixation. Thus, endosymbiotic gene transfer

can be advantageous in the absence of any additional benefit to the host cell, providing new insight into the processes that have shaped eukaryotic genome evolution.

Main

Endosymbiosis has underpinned two of the most important innovations in the history of life on Earth (Archibald 2015a; Martin, et al. 2015). The endosymbiosis of the alphaproteobacterium that became the mitochondrion led to the emergence and radiation of the eukaryotes (Yang, et al. 1985; Martin and Müller 1998; Roger, et al. 2017), and the endosymbiosis of the cyanobacterium that became the chloroplast first enabled oxygenic photosynthesis in eukaryotes (Martin and Kowallik 1999; Archibald 2015b). The function and evolution of both organelles is inextricably linked with energy metabolism and the evolution of the eukaryotic cell (Lane and Martin 2010; Lane 2014; Booth and Doolittle 2015a, b; Lane and Martin 2015; Lynch and Marinov 2017; Roger, et al. 2017; Lynch and Marinov 2018), and has given rise to the multicellular organisms that dominate the biosphere (Bar-On, et al. 2018).

Following the endosymbioses of the bacterial progenitors of mitochondrion and the chloroplast there was a dramatic reduction in the gene content of the endosymbiont genomes such that they harbour fewer than 5% of the genes found in their free-living prokaryotic relatives (Gray, et al. 1999; Timmis, et al. 2004; Green 2011). While many of the original endosymbiont genes have been lost (Lynch, et al. 2006; McCutcheon and Moran 2012; Smith and Keeling 2015; Smith 2016), others have been transferred to the host nuclear genome and their products imported back into the organelle where they function in a process known as endosymbiotic gene transfer (Martin, et al. 2002; Brown 2003; Deusch, et al. 2008; Thiergart, et al. 2012; Dagan, et al. 2013). For example, the mitochondrion of humans (Calvo and Mootha 2010) and chloroplasts of plants (Ferro, et al. 2010) 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. Moreover, the process of 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 (McCutcheon and Moran 2012; Husnik, et al. 2013) and the endosymbiosis of the cyanobacterium that became the chromatophore in *Paulinella* (Nakayama and Ishida 2009; Nowack, et al. 2010; Reyes-Prieto, et al. 2010; Singer, et al. 2017; Nowack and

Weber 2018). Thus, endosymbiont genome reduction and endosymbiotic gene transfer are recurring themes in the evolution of eukaryotic nuclear and cytoplasmic genomes.

Given, the importance of endosymbiotic gene transfer to the evolution of eukaryotic genomes, several hypotheses have been proposed to explain why it occurs (Herrmann 1997; Martin and Herrmann 1998; Daley and Whelan 2005; Reyes-Prieto, et al. 2006; Speijer, et al. 2020). For example, it has been proposed that it protects endosymbiont genes from mutational hazard (Allen and Raven 1996; Lynch, et al. 2006; Smith 2016; Speijer, et al. 2020), and that it enables endosymbiont genes that are otherwise trapped in a haploid genome to recombine and thus escape from Muller's ratchet (Muller 1964; Lynch 1996; Martin and Herrmann 1998; Lynch, et al. 2006; Neiman and Taylor 2009; Smith 2016). It has also been proposed that endosymbiotic gene transfer is an inevitable consequence of a constant stream of endosymbiont genes entering the nucleus (Doolittle 1998), and that transfer to the nuclear genome allows the host cell to gain better control over the replication and function of the organelle (Herrmann 1997) allowing better cellular network integration (Nowack, et al. 2010; Reyes-Prieto 2015). However, mutation rates of organellar genes are often not higher than nuclear genes (Wolfe, et al. 1987; Lynch, et al. 2006; Lynch, et al. 2007; Drouin, et al. 2008; Smith 2015; Smith and Keeling 2015; Smith 2016; Grisdale, et al. 2019) 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 (Lynch 1996; Neiman and Taylor 2009) chloroplasts appear largely to escape this effect (Wolfe, et al. 1987; Lynch 1997) likely due to gene conversion (Khakhlova and Bock 2006), 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, and may simply be a projection of anthropocentric ideals of centralised control onto the nucleus of the host cell. 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.

We hypothesised that an advantage for endosymbiotic gene transfer 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 harbours multiple copies of the organellar genome (Bendich 1987; Cole 2016). 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 (Shokolenko, et al. 2009) and to enable the organelle to achieve high protein abundance for genes encoded in the organellar genome (Bendich 1987). Thus, while a typical diploid eukaryotic cell contains two copies of the nuclear genome, the same cell contains hundreds to hundreds of thousands of copies of its organellar genomes (Bendich 1987; Cole 2016). As DNA costs energy and cellular resources to biosynthesise (Lynch and Marinov 2015), the cost to the cell of encoding a gene in the organellar and nuclear genome is different. To quantify this difference, we evaluated the cost of encoding a gene in the nuclear or organellar genome. Here, the cost of a gene was considered to be the cost of the chromosome divided by the number of genes on that chromosome to account for introns, structural, and regulatory elements (we also included the cost of the requisite number of histone proteins contained in nucleosomes for nuclear genes). This revealed that the 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 (Figure 1A). This difference is further enhanced if the biosynthesis cost of just the coding sequences of the genes are compared directly (Figure 1B). Thus, the cost to the cell of encoding a gene in the organellar genome is substantially higher than the cost of encoding the same gene in the nuclear genome. Consequently, for any essential organellar gene the cell may be able to save resources by transferring that gene from the organellar genome to the nuclear genome. 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 Methods for sources of genome copy numbers). We hypothesised that if the energy saved by transferring a gene offset the cost of importing the required abundance of gene product back into the organelle then this would provide a direct energetic and fitness advantage to the host cell for endosymbiotic gene transfer.

To test this hypothesis, we assessed the conditions under which it is more energetically favourable to encode a gene in the organellar or nuclear genome. Here, the free energy of endosymbiotic gene transfer (which we define as the difference in energy cost between a cell which encodes a given gene in the organellar genome and a cell which encodes the same gene in the nuclear genome and imports the requisite amount of gene product into to the organelle, see Methods) was computed for an average length bacterial gene as a function of protein abundance, protein import cost, and organellar genome copy number. This revealed that there is a simple relationship such that the higher the copy number of the organellar genome, the more energy that is liberated by endosymbiotic gene transfer and thus the more protein that can be imported into the organelle while still reducing the overall energetic cost of the cell (Figure 2A). To simulate the organellar genome reduction that would result if all such energetically favourable endosymbiotic gene transfers occurred, the complete genomes with measured protein abundances for an alphaproteobacterium (*Bartonella henselae*) and a cyanobacterium (*Microcystis aeruginosa*) were subject to a simulated endosymbiosis. Here, a range of host cell sizes was simulated such that they encompassed the majority of diversity exhibited by extant eukaryotes (Milo 2013) and would thus likely encompass the size range of the host cell that originally engulfed the bacterial organellar progenitors. This range extended from a small unicellular yeast-like cell (10^7 proteins), to a typical unicellular algal cell (10^8 proteins) to a large metazoan/plant cell (10^9 proteins). Each of these cell types were then considered to allocate a realistic range of total cellular protein to mitochondria/chloroplasts representative of extant eukaryotic cells (Supplemental Table S1). For each simulated endosymbiosis, the free energy of endosymbiotic gene transfer was calculated for each gene given its measured protein abundance (Wang, et al. 2015) and a realistic range of protein import costs (including the total biosynthetic cost of the protein import machinery, See Methods). 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 favourable to the cell to transfer the majority of organellar genes to the nuclear genome and re-import the proteins back to the organelle (Figure 2B and 2C). Here, only the proteins with the highest abundance, and thus which occur the largest import cost, are retained in the organellar genomes. This phenomenon holds true

even if extreme costs for protein import ten times those that have been measured are considered (Supplemental Figure S1), or if organellar protein import is inefficient (or organellar protein turn-over is higher than cytosolic protein turn-over) such that 50% of cytosolic synthesised protein fails to be imported (or is turned over) and is thus wasted (Supplemental Figure S2). Thus, it is more energy efficient for the cell to transfer the majority of organellar genes to the nuclear genome and import the proteins into the organelle.

To estimate the strength of selection that would act on the change in energy incurred from an endosymbiotic gene transfer event, the free energy of endosymbiotic gene transfer for each gene was placed in context 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 genome copy number, and protein import cost that is representative of a broad range of eukaryotic cells (Figure 3A and B, Supplemental Figures S3 – S8, Supplemental Table S2). This revealed that for even modest per-cell endosymbiont genome copy numbers (≥100 copies per cell) the selection coefficients for the transfer of the majority of endosymbiont genes are relatively large ~$1x10^{-4}$ (Figure 3, Supplemental Figures S3 – S8), ~10,000 times stronger than the selection coefficient acting against disfavoured synonymous codons (Hartl, et al. 1994). Moreover, for high per-cell endosymbiont genome copy numbers (≥1000 genome copies per cell) these selection coefficients are large (~1×10^{-3}) and similar to the strength of selection that caused the allele conferring lactose tolerance to rapidly sweep through human populations in ~500 generations (Bersaglieri, et al. 2004). In contrast, selection coefficients for retention of genes in the organellar genome only occur when organellar genome copy numbers are low, and/or when large proportions of cellular resources are invested in organelle (Figure 3A and B, Supplemental Figures S3 – S8). However, with the exception of very highly abundant proteins (discussed below) these selection coefficients are generally weaker. Thus, over a broad range of host cell sizes, organellar genome copy numbers, organellar fractions, and per-protein ATP import costs, endosymbiotic gene transfer of the majority of genes is sufficiently energetically advantageous that any such transfer events, if they occurred, would rapidly reach fixation (Supplemental Figure S9). Thus, endosymbiotic gene transfer of the majority of genes is intrinsically advantageous to the cell in the absence of any other additional benefits.

Although the energetic advantage is sufficiently high to explain why such events would become fixed if they occurred, it is not proposed that it is the only factor that influences the location of an organellar gene. 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 rachet, enhanced nuclear control, and drift will act antagonistically or synergistically with the free energy of endosymbiotic gene transfer to influence the set of genes that are retained in, or transferred from, the organellar genomes. Moreover, the free energy of endosymbiotic gene transfer provides a mechanistic basis for selection to act with or against Doolittle's "You are what you eat" ratchet for endosymbiotic gene transfer (Doolittle 1998). 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 catalysing 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 is predominantly thought to be mediated by mutation pressure and drift (Lynch, et al. 2006), the elevated per-cell endosymbiont genome copy number also provides a substantial energetic reward to the host cell for complete gene loss. Thus, the high genome copy number required to protect DNA from damage (Shokolenko, et al. 2009) and facilitate high levels of protein production (Bendich 1987), also provides the energetic incentive for the cell to delete endosymbiont genes as well as transfer them to the nuclear genome.

The analysis presented here shows that for a broad range of cell sizes and resource allocations that endosymbiotic gene transfer of the majority of organellar genes is energetically favourable and thus advantageous to the cell. However, it also showed that retention of genes in the organellar genomes is energetically favourable under conditions where the encoded organellar protein is required in very high abundance and/or the copy number of the organellar genome is low (Figure 2B, 2C, 3A and 3B). For example, in large plant cells such as those in the leaves of *Arabidopsis thaliana* it is unfavourable to transfer the *rbcL* gene encoding the RuBisCO large subunit from the chloroplast genome to the nuclear genome, as although it would save 8.7 $\times 10^{7}$ ATP per cell in DNA biosynthesis costs it would incur a daily cost of ~3.96 $\times 10^{12}$ ATP per cell (0.17% of the daily energy budget of the cell) just to import the required amount of RuBisCO large subunit back into the

chloroplast (see methods). Consequently, from a cost perspective it is energetically favourable to the cell to retain this gene in the chloroplast genome. The same is also true for 62 of the 88 genes currently found in the chloroplast genome in *Arabidopsis thaliana* (Supplemental Table S3) such that selection would act against transfer of these genes from the chloroplast genome. In, contrast it is energetically favourable to transfer the majority of genes from the mitochondrial genome to the nuclear genome in *Arabidopsis* (99 out of 122), and all of the genes encoded in the human mitochondrial genome to the human nuclear genome (Supplemental Table S3). Thus, requirement for high protein abundance and low genome copy numbers will act synergistically with factors such as the requirement for redox regulation of gene expression (Allen 2015) to favour retention of genes in the organellar genomes.

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 (Supplemental Figure S10). 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.

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 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 (Rogozin, et al. 2012) means that it is more difficult for endosymbiotic gene transfer to operate in the reverse direction (i.e. from the nucleus to

organelle). Collectively, this would create a 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 evolution. Thus, current organellar and nuclear gene contents predominantly reflect past pressures to transfer genes to the nuclear genome.

Endosymbiotic gene transfer is a recurring theme in the evolution of the eukaryotic tree of life. The discovery that the free energy of endosymbiotic gene transfer can provide an energetic advantage to the cell for 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. Moreover, it reveals that organelles have surrendered the vast majority of their genes to the nuclear genome for the sake of the greater good of the cell.

Materials and Methods

Data sources

The *Arabidopsis thaliana* genome sequence and corresponding set of representative gene models were downloaded from Phytozome V13 (Goodstein, et al. 2012). 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, the *Microcystis aeruginosa* NIES-843 genome sequence and gene models from assembly version ASM1062v1 were each downloaded from Ensembl (Yates, et al. 2020). The *Saccharomyces cerevisiae* sequence and gene models from assembly version R64-2-1_20150113 were downloaded from the *Saccharomyces* Genome Database (Cherry, et al. 2012). Protein abundance data for all species were obtained from PAXdb v4.1 (Wang, et al. 2015).

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 (Chen, et al. 2016) and (Lynch and Marinov 2015) and are provided in Supplemental Table S4. *The Homo sapiens* mitochondrial genome copy number of 5000 was obtained from (Cole 2016). The *Saccharomyces cerevisiae* mitochondrial genome copy number of 200 was obtained from (Miyakawa 2017). The *Arabidopsis thaliana* chloroplast genome copy number of 1500 was obtained from (Zoschke, et al.

2007) and the *Arabidopsis thaliana* mitochondrial genome copy number of 100 was obtained from (Cole 2016).

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 180bp (147bp for the two turns of DNA around the histone octamer and 33bp for the spacer) (Lynch and Marinov 2015). For organellar chromosomes there are no histones/nucleosomes and thus the biosynthetic cost of genes in organellar chromosomes was calculated as cost of the DNA divided by the number of genes on the chromosome (Supplemental Table S5).

The average gene length used for the simulation study in Figure 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 (Soll and Schleiff 2004; Jarvis 2008; Wiedemann and Pfanner 2017) they share many commonalities including the requirement for energy in the form of nucleoside triphosphate hydrolysis (Schatz and Dobberstein 1996). 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 (Mokranjac and Neupert 2008; Shi and Theg 2013; Backes and Herrmann 2017). 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, Supplemental Table S6) 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 (Supplemental

Table S7). 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 (Supplemental Table S7). In all cases the proteins that were predicted to be imported into the organelle were identified using TargetP-2.0 (Almagro Armenteros, et al. 2019) and protein abundance was calculated using measured protein abundance estimates for each species obtained from PAXdb 4.0 (Wang, et al. 2015), assuming a total cell protein content of $1x10^9$ proteins for a human cell, $1x10^7$ proteins for a yeast cell and 2.5×10^{10} proteins for an *Arabidopsis thaliana* cell. As we modelled 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 (Almagro Armenteros, et al. 2019) and protein abundance estimates obtained from PAXdb 4.0 (Wang, et al. 2015). The proportion of cellular resources are provided in Supplemental Table S1 and were used to provide the indicative regions or parameter space occupied by metazoa, yeast and plants shown on Figure 2B and 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 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

$$\Delta P = P_{end} - P_{cyt} - P_{import} \quad [3]$$

The energetic cost of producing a protein in the endosymbiont and in the cytosol are assumed to be equal and thus

$$\Delta P = P_{import} \quad [4]$$

It should be noted here that although the P_{end} and P_{cyt} are assumed to be equal for the majority of calculations an analysis was conducted wherein an inefficient protein import system was assumed such that 50% of protein failed to be imported and thus must be turned over (Supplemental Figure S2). Even under these conditions it is still energetically favourable to encode organellar genes in the nuclear genome for realistic estimates of cell sizes and investment in organelles.

P_{import} is evaluated as the product of 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}), 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 (Mokranjac and Neupert 2008; Shi and Theg 2013; Backes and Herrmann 2017). For the purposes of this study we used these measured ranges and also modelled 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

$$D_{end} = A_{DNA} C_{end} \quad [6]$$

And

$$D_{nuc} = A_{DNA} C_{nuc} \quad [7]$$

Such that

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

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 are 40.55 ATP and 40.14 ATP respectively (Chen, et al. 2016). Thus

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

Where positive values of ΔE_{EGT} correspond to genes for which it is more energetically favourable to be encoded in the nuclear genome, and negative values correspond to genes for which it is more energetically favourable to be encoded in the endosymbiont genome.

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 sever 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 (Milo 2013). 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. ~2% for yeast (Uchida, et al. 2011), ~20% for metazoan cells (David 1977) and ~50% of the non-vacuolar volume of plant cells (Winter, et al. 1994)). 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 (Wang, et al. 2015) for a realistic range of endosymbiont genome copy numbers using equation 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 respectively). The number of genes

where ΔE_{EGT} was positive was recorded as these genes comprise the cohort that are energetically favourable to be encoded in the nuclear genome. All calculated values for ΔE_{EGT} for both the model organisms are provided in Supplemental Table S2.

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 x 10^7 proteins per cell and 1 x 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-hour doubling time such that all genomes and proteins are produced in the required abundance every 24-hour period. As previously defined (Lynch and Marinov 2015), the energy required for cell growth was modelled 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 $cell^{-1}$ $hour^{-1}$, and V is in units of μm^3 (Lynch and Marinov 2015). 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-hour life cycle.

$$\mathrm{E}_{A/D} = \frac{\Delta E_{EGT}}{E_R} \quad [13]$$

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

$$s = \mathrm{E}_{A/D} \quad [14]$$

As ΔE_{EGT} can be positive or negative as described above, s is therefore also positive or negative depending on 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 zero 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}). All calculated values for *s* for both the model alphaproteobacterium (*Bartonella henselae*) and cyanobacterium (*Microcystis aeruginosa*) are provided in Supplemental Table S1.

It should be noted here that variation in the doubling time will have a direct effect on the estimate of the selection coefficients. Decreasing cell doubling time 10-fold increases the values of *s* by a factor of 10 and *vice versa*, such that cells with shorter doubling times would experience stronger selection on the free energy of endosymbiotic gene transfer and cells with longer doubling times would experience weaker selection.

Similarly, the effect of protein turnover was not included as estimates for protein turn over were not available for each protein considered in these analyses. However, the effect of protein turnover is to increase the total amount of protein that must be produced within the life cycle of the cell. Thus, for the purposes of this analysis can be considered equivalent to increasing the cellular investment in organelles.

Estimating time to fixation

Fixation times for endosymbiotic gene transfer events for a range of observed selection coefficients from 1 x 10^{-5} to 1 x 10^{-2} were estimated using a Wright–Fisher model with selection and drift (Fisher 1930; Wright 1931) implemented in a simple evolutionary dynamics simulation (Niklaus and Kelly 2018). The effective population size for these simulations was set as 1 x 10^{7}, as is

representative of unicellular eukaryotes (Lynch and Conery 2003) and multicellularity in eukaryotes is not thought to have evolved until after the endosymbiosis of either the mitochondrion or the chloroplast.

The cost of transferring the rbcL gene encoding RuBisCO large subunit from the chloroplast to the nuclear genome in Arabidopsis thaliana

The total number of proteins contained in an *Arabidopsis thaliana* leaf cell is 2.5 x 10^{10} proteins (Heinemann, et al. 2020). The fraction of cellular protein that is invested in RuBisCO large subunit (F_{rbcL}) is 0.165 (Li, et al. 2017) and thus the number of RuBisCO large subunit proteins per cell (N_p) is estimated to be 4.13 x 10^9. The cost of import (P_{import}) of a protein to the chloroplast is 2 ATP per amino acid residue (Shi and Theg 2013). The length of the polypeptide (L_{prot}) comprising the RuBisCO large subunit is 480 amino acids (1440 nucleotides). The ATP biosynthesis cost of a single copy of the *rbcL* gene in double stranded DNA (A_{DNA}) is 58132 ATP. The copy number of the chloroplast genome in a typical *Arabidopsis thaliana* leaf cell (C_{end}) is 1500 copies (Zoschke, et al. 2007). *Arabidopsis thaliana* is diploid and thus the copy number of the nuclear genome (C_{nuc}) is 2. Thus, using equation 8 above the ATP that would be saved by transferring the DNA encoding the *rbcL* gene from the chloroplast genome to the nuclear genome is evaluated as

$$58132\ (1500 - 2) = 8.7 \text{ x } 10^7\ ATP \quad [15]$$

Using equation 5 above the ATP that would be required to import the RuBisCO large subunit into the chloroplast is thus evaluated as

$$480 \text{ x } 2 \text{ x } 4.13 \text{ x } 10^9 = 3.96 \text{ x } 10^{12}\ ATP \quad [16]$$

And thus

$$\Delta E_{EGT} = 8.7 \text{ x } 10^7 - 3.96 \text{ x } 10^{12} = -3.96 \text{ x } 10^{12}\ ATP \quad [17]$$

Given that an *Arabidopsis thaliana* leaf mesophyll cell has a volume of ~49 µm^3 (Ramonell, et al. 2001) the energy consumption rate cell was calculated using equation 10 be consumes 5.2 x 10^{12} ATP hour^{-1}. Assuming an experimentally determined *in vivo* degradation rate for RuBisCO large subunit of K_D = 0.052 (Li, et al. 2017), the recurring daily cost of importing new RuBisCO large subunits into the chloroplast is evaluated as

$$L_{prot} C_{import} N_p K_D = 2.1\ x\ 10^{11}\ ATP \quad [18]$$

Thus, if the *rbcL* gene was transferred from the chloroplast genome to the nuclear genome in *Arabidopsis thaliana*, the daily cost of importing the RuBisCO large subunit back into the chloroplast would consume ~0.17% of the total operational energy budget of the cell.

Acknowledgements

This work was funded by the Royal Society and the European Union's Horizon 2020 research and innovation program under grant agreement number 637765. The author would like to thank Thomas A. Richards and John M. Archibald for their comments on the manuscript.

Author Contributions

SK conceived study, conducted the analysis, and wrote the manuscript.

Figures

Figure 1

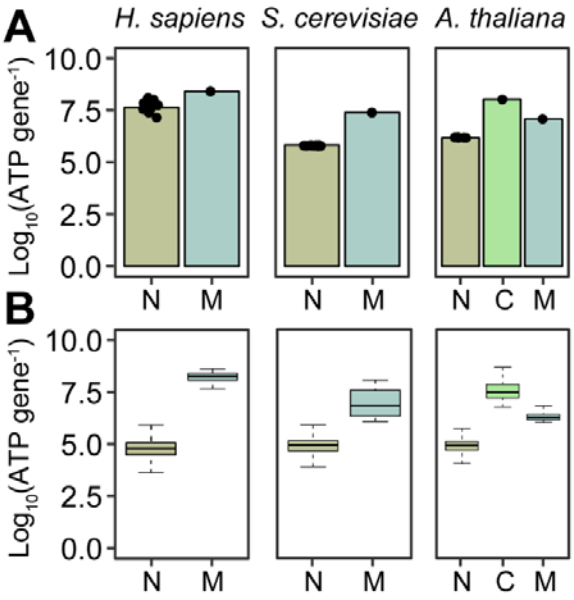

Figure 1. The per-cell biosynthetic cost of nuclear and organellar genes in three representative eukaryotes. **A**) The ATP biosynthesis costs of nuclear (N), chloroplast (C), and mitochondrial (M) genes calculated as the cost of the chromosome divided by the number of genes contained within that chromosome. Nuclear chromosomes include the cost of nucleosomes, organellar chromosomes only included the cost of the DNA. In the case of the nuclear genes the height of bar depicts the mean cost of all nuclear chromosomes with individual points showing all chromosomes overlaid on top the bar plots. **B**) The ATP biosynthesis cost of just the coding sequences of the genes. In both A and B, the costs were computed assuming a diploid nuclear genome, a per-cell mitochondrial genome copy number of 5000, 200 and 100 for the in *H. sapiens*, *S. cerevisiae* and *A. thaliana*, respectively, and a per cell chloroplast genome copy number of 1500 in *A. thaliana*.

Figure 2

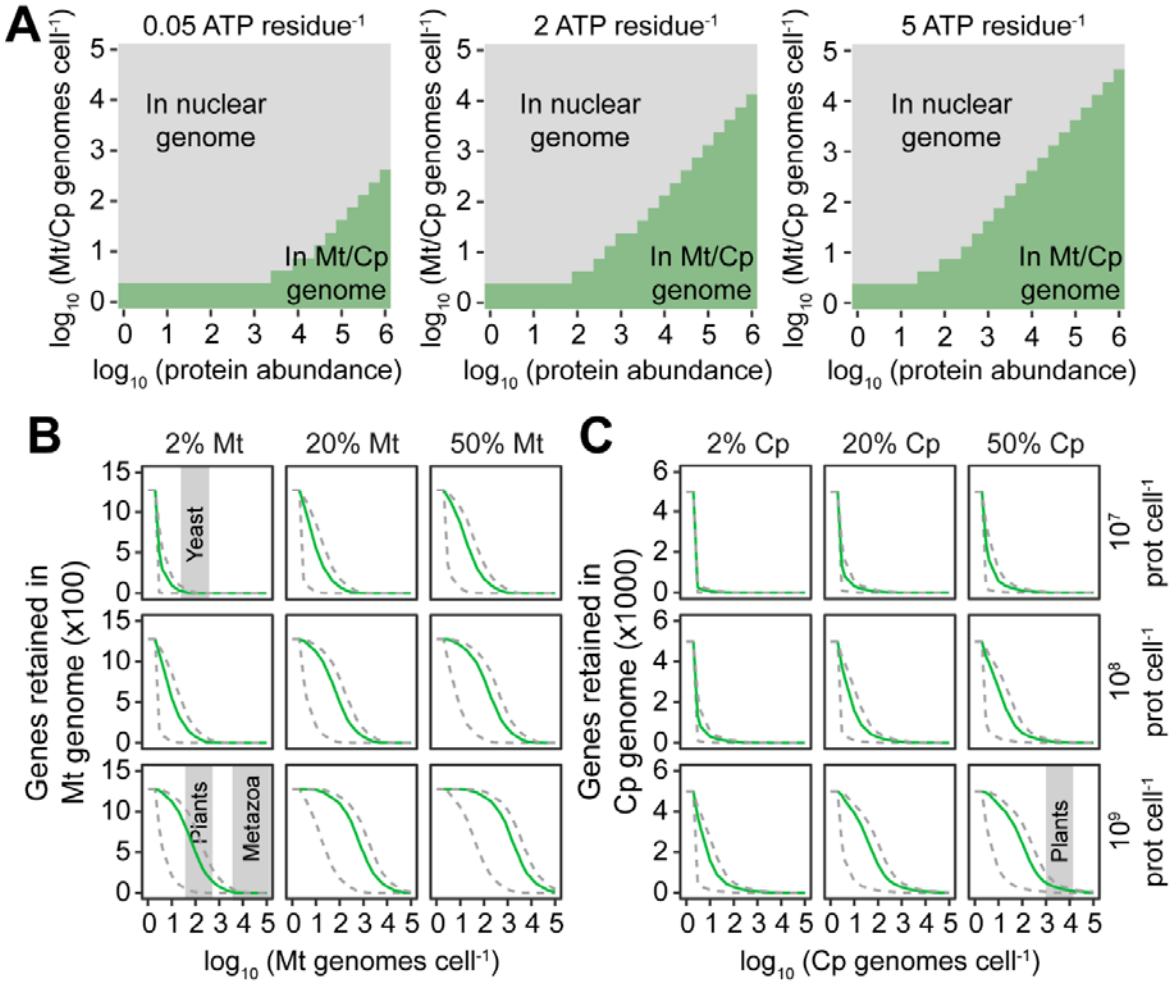

Figure 2. The minimum cost location to the cell of organellar genes encoding an organellar localised protein. **A**) The minimum cost location of an organellar gene for a range of per-protein import costs, organellar genome copy numbers, and encoded protein abundance. The modelled per-residue protein import cost is shown above each plot. The grey shaded fractions of the plots indicate the regions of parameter space where it is more energetically favourable to the cell to encode an organellar gene in the nuclear genome and import the requisite amount of protein. The green shaded fractions of the plots indicate the regions of parameter space where it is more energetically favourable to the cell to encode the gene in the organellar genome. **B**) The number of genes in the alphaproteobacterial (mitochondrial) genome for which it is more energetically favourable to the cell for the gene to be retained in the organellar genome. Green lines assume a per-residue protein import cost of 2 ATP per amino acid. Grey dashed lines indicate lower and upper cost bounds of 0.05 ATP and 5 ATP per residue respectively. **C**) As in B but for the

cyanobacterial (chloroplast) genome. Grey shaded areas on plots are provided for illustrative purposes to indicate the regions of parameter space occupied by yeast, metazoan and plant cells. Cp: chloroplast. Mt: mitochondrion.

Figure 3

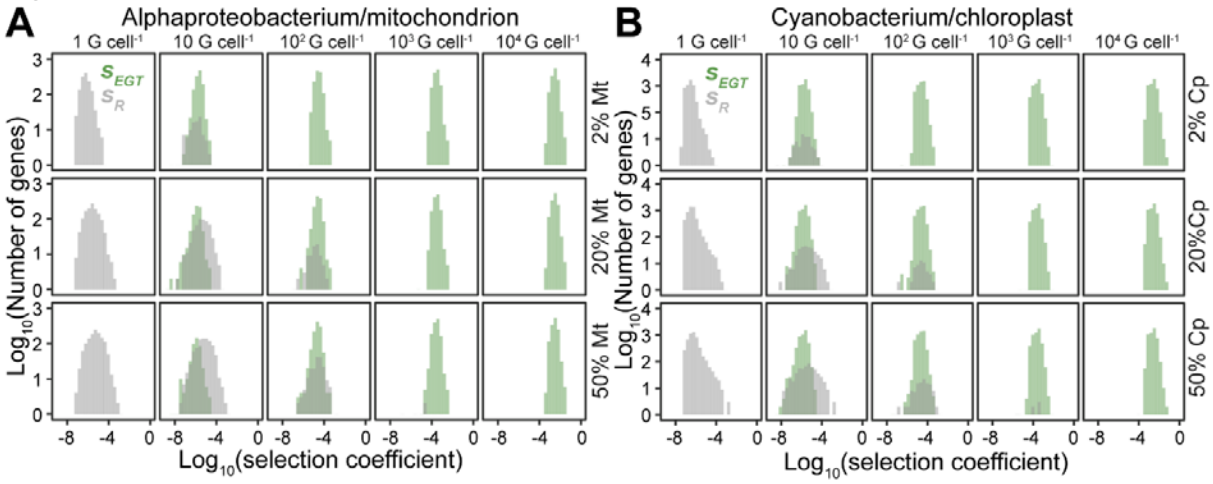

Figure 3. Selection coefficients for retention (S_R, grey) or endosymbiotic gene transfer (S_{EGT}, green) of all genes encoded in the example alphaproteobacterial and cyanobacterial genomes. Coefficients were computed accounting for protein abundance, host cell organellar fraction, organellar genome copy number per cell, and host cell energy consumption. Plots shown are for a simulated host cell comprising 1 x 10^7 proteins and a protein import cost of 2 ATP per residue, plots for other host cell protein contents and protein import costs are provided in Supplemental Figures S3-S8. **A**) Selection coefficients of all genes encoded in the alphaproteobacterium genome. **B**) Selection coefficients for all genes encoded in the cyanobacterial genome. S_R and S_{EGT} have opposite signs (see methods). To simplify the display and enable direct comparison, the absolute value of the selection coefficients of each gene are plotted and green shading is used to indicate genes in the S_{EGT} fraction and grey shading indicates genes int the S_R fraction of the genome. Mt, mitochondrion. Cp, chloroplast. G, genomes.

References

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A 112:10231-10238.

Allen JF, Raven JA. 1996. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J Mol Evol 42:482-492.

Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance 2.

Archibald John M. 2015a. Endosymbiosis and Eukaryotic Cell Evolution. Current Biology 25:R911-R921.

Archibald JM. 2015b. Genomic perspectives on the birth and spread of plastids. Proceedings of the National Academy of Sciences 112:10147-10153.

Backes S, Herrmann JM. 2017. Protein Translocation into the Intermembrane Space and Matrix of Mitochondria: Mechanisms and Driving Forces. Frontiers in Molecular Biosciences 4.

Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on Earth. Proceedings of the National Academy of Sciences 115:6506-6511.

Bendich AJ. 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? Bioessays 6:279-282.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111-1120.

Booth A, Doolittle WF. 2015a. Eukaryogenesis, how special really? Proceedings of the National Academy of Sciences 112:10278-10285.

Booth A, Doolittle WF. 2015b. Reply to Lane and Martin: Being and becoming eukaryotes. Proceedings of the National Academy of Sciences 112:E4824-E4824.

Brown JR. 2003. Ancient horizontal gene transfer. Nature Reviews Genetics 4:121-132.

Calvo SE, Mootha VK. 2010. The mitochondrial proteome and human disease. Annu Rev Genomics Hum Genet 11:25-44.

Chen W-H, Lu G, Bork P, Hu S, Lercher MJ. 2016. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nature Communications 7:11334.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40:D700-705.

Cole LW. 2016. The Evolution of Per-cell Organelle Number. Front Cell Dev Biol 4:85.

Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, et al. 2013. Genomes of

Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol 5:31-44.
Daley DO, Whelan J. 2005. Why genes persist in organelle genomes. Genome Biology 6:110.
David H. 1977. Quantitative Ultrastructural Data of Animal and Human Cells: Gustav Fischer.
Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor. Molecular Biology and Evolution 25:748-761.
Doolittle WF. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14:307-311.
Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol 49:827-831.
Ferro M, Brugière S, Salvi D, Seigneurin-Berny D, Court M, Moyet L, Ramus C, Miras S, Mellal M, Le Gall S, et al. 2010. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. Mol Cell Proteomics 9:1063-1084.
Fisher RAS. 1930. The genetical theory of natural selection. Oxford: Clarendon Press.
Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178-1186.
Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. Science 283:1476-1481.
Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J 66:34-44.
Grisdale CJ, Smith DR, Archibald JM. 2019. Relative Mutation Rates in Nucleomorph-Bearing Algae. Genome Biology and Evolution 11:1045-1053.
Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias. Genetics 138:227-234.
Heinemann B, Künzler P, Braun H-P, Hildebrandt TM. 2020. Estimating the number of protein molecules in a plant cell: a quantitative perspective on proteostasis and amino acid homeostasis during progressive drought stress. bioRxiv:2020.2003.2017.995613.
Herrmann R. 1997. Eukaryotism, towards a new interpretation. In. Eukaryotism and symbiosis: Springer. p. 73-118.
Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson AC, et al. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153:1567-1578.
Jarvis P. 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. New Phytologist 179:257-285.

Khakhlova O, Bock R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. The Plant Journal 46:85-94.
Lane N. 2014. Bioenergetic constraints on the evolution of complex life. Cold Spring Harb Perspect Biol 6:a015982.
Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929-934.
Lane N, Martin WF. 2015. Eukaryotes really are special, and mitochondria are why. Proceedings of the National Academy of Sciences 112:E4823-E4823.
Li L, Nelson CJ, Trösch J, Castleden I, Huang S, Millar AH. 2017. Protein Degradation Rate in Arabidopsis thaliana Leaf Growth and Development. The Plant Cell 29:207-228.
Lynch M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. Mol Biol Evol 14:914-925.
Lynch M. 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. Mol Biol Evol 13:209-220.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401-1404.
Lynch M, Koskella B, Schaack S. 2006. Mutation Pressure and the Evolution of Organelle Genomic Architecture. Science 311:1727-1730.
Lynch M, Lynch PSTSM, Walsh B. 2007. The Origins of Genome Architecture: Oxford University Press, Incorporated.
Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. Proceedings of the National Academy of Sciences 112:15690-15695.
Lynch M, Marinov GK. 2017. Membranes, energetics, and evolution across the prokaryote-eukaryote divide. eLife 6:e20437.
Lynch M, Marinov GK. 2018. Response to Martin and colleagues: mitochondria do not boost the bioenergetic capacity of eukaryotic cells. Biology Direct 13:26.
Martin W, Herrmann RG. 1998. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? Plant Physiology 118:9-17.
Martin W, Kowallik K. 1999. Annotated English translation of Mereschkowsky's 1905 paper ‘Über Natur und Ursprung der Chromatophoren imPflanzenreiche’. European Journal of Phycology 34:287-295.
Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37-41.
Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. 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 99:12246-12251.
Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. Philos Trans R Soc Lond B Biol Sci 370:20140330.
McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology 10:13-26.

Milo R. 2013. What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays 35:1050-1055.
Miyakawa I. 2017. Organization and dynamics of yeast mitochondrial nucleoids. Proc Jpn Acad Ser B Phys Biol Sci 93:339-359.
Mokranjac D, Neupert W. 2008. Energetics of protein translocation into mitochondria. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1777:758-762.
Muller HJ. 1964. The relation of recombination to mutational advance. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 1:2-9.
Nakayama T, Ishida K. 2009. Another acquisition of a primary photosynthetic organelle is underway in Paulinella chromatophora. Curr Biol 19:R284-285.
Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes. Proceedings of the Royal Society B: Biological Sciences 276:1201-1209.
Niklaus M, Kelly S. 2018. The molecular evolution of C4 photosynthesis: opportunities for understanding and improving the world's most productive plants. Journal of Experimental Botany 70:795-804.
Nowack ECM, Vogel H, Groth M, Grossman AR, Melkonian M, Glöckner G. 2010. Endosymbiotic Gene Transfer and Transcriptional Regulation of Transferred Genes in Paulinella chromatophora. Molecular Biology and Evolution 28:407-422.
Nowack ECM, Weber APM. 2018. Genomics-Informed Insights into Endosymbiotic Organelle Evolution in Photosynthetic Eukaryotes. Annual Review of Plant Biology 69:51-84.
Ramonell KM, Kuang A, Porterfield DM, Crispi ML, Xiao Y, McClure G, Musgrave ME. 2001. Influence of atmospheric oxygen on leaf structure and starch deposition in Arabidopsis thaliana. Plant Cell Environ 24:419-428.
Reyes-Prieto A. 2015. The basic genetic toolkit to move in with your photosynthetic partner. Frontiers in Ecology and Evolution 3.
Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial Contribution to Algal Nuclear Genomes Is Primarily Limited to Plastid Functions. Current Biology 16:2320-2325.
Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, Boo SM, Nakayama T, Ishida K, Bhattacharya D. 2010. Differential gene retention in plastids of common recent origin. Mol Biol Evol 27:1530-1537.
Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The Origin and Diversification of Mitochondria. Curr Biol 27:R1177-r1192.
Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. Biol Direct 7:11.
Schatz G, Dobberstein B. 1996. Common Principles of Protein Translocation Across Membranes. Science 271:1519-1526.
Shi L-X, Theg SM. 2013. Energetic cost of protein import across the envelope membranes of chloroplasts. Proceedings of the National Academy of Sciences 110:930-935.

Shokolenko I, Venediktova N, Bochkareva A, Wilson GL, Alexeyev MF. 2009. Oxidative stress induces degradation of mitochondrial DNA. Nucleic Acids Res 37:2539-2548.
Singer A, Poschmann G, Mühlich C, Valadez-Cano C, Hänsch S, Hüren V, Rensing SA, Stühler K, Nowack ECM. 2017. Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba Paulinella chromatophora. Curr Biol 27:2763-2773.e2765.
Smith DR. 2015. Mutation rates in plastid genomes: they are lower than you might think. Genome Biol Evol 7:1227-1234.
Smith DR. 2016. The mutational hazard hypothesis of organelle genome evolution: 10 years on. Molecular Ecology 25:3769-3775.
Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proceedings of the National Academy of Sciences 112:10177-10184.
Soll J, Schleiff E. 2004. Protein import into chloroplasts. Nat Rev Mol Cell Biol 5:198-208.
Speijer D, Hammond M, Lukeš J. 2020. Comparing Early Eukaryotic Integration of Mitochondria and Chloroplasts in the Light of Internal ROS Challenges: Timing is of the Essence. mBio 11:e00955-00920.
Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol 4:466-485.
Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123-135.
Uchida M, Sun Y, McDermott G, Knoechel C, Le Gros MA, Parkinson D, Drubin DG, Larabell CA. 2011. Quantitative analysis of yeast internal architecture using soft X-ray tomography. Yeast 28:227-236.
Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163-3168.
Wiedemann N, Pfanner N. 2017. Mitochondrial Machineries for Protein Import and Assembly. Annual Review of Biochemistry 86:685-714.
Winter H, Robinson DG, Heldt HW. 1994. Subcellular volumes and metabolite concentrations in spinach leaves. Planta 193:530-535.
Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences 84:9054-9058.
Wright S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.
Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. 1985. Mitochondrial origins. Proceedings of the National Academy of Sciences 82:4443-4447.
Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. Nucleic Acids Res 48:D682-d688.

Zoschke R, Liere K, Börner T. 2007. From seedling to mature plant: Arabidopsis plastidial genome copy number, RNA accumulation and transcription are differentially regulated during leaf development. The Plant Journal 50:710-722.