Connecting species-specific extents of genome reduction in mitochondria and plastids

Konstantinos Giannakis[1], Luke Richards[2], Kazeem A. Dauda[1], Iain G. Johnston[1,3,*]

1. Department of Mathematics, University of Bergen, Bergen, Norway
2. School of Life Sciences, University of Warwick, Coventry, UK
3. Computational Biology Unit, University of Bergen, Bergen, Norway
* correspondence to iain.johnston@uib.no

Abstract

Mitochondria and plastids have both dramatically reduced their genomes since the endosymbiotic events that created them. The similarities and differences in the evolution of the two organelle genome types has been the target of discussion and investigation for decades. Ongoing work has suggested that similar mechanisms may modulate the reductive evolution of the two organelles in a given species, but quantitative data and statistical analyses exploring this picture remain limited outside of some specific cases like parasitism. Here, we use cross-eukaryote organelle genome data to explore evidence for coevolution of mitochondrial and chloroplast genome reduction. Controlling for differences between clades and pseudoreplication due to relatedness, we find that mtDNA and ptDNA gene retention are related across taxa, in a generally positive correlation that appears to differ quantitatively across eukaryotes, for example, between algal and non-algal species. We find limited evidence for coevolution of specific mtDNA and ptDNA gene pairs, suggesting that the similarities between the two organelle types may be due mainly to independent responses to consistent evolutionary drivers.

Introduction

Mitochondria and plastids (the class of organelles to which chloroplasts belong) are essential compartments in eukaryotic cells. Originally independent organisms, both organelle classes arose from endosymbiotic capture and subsequent evolution (Zimorski et al. 2014; Sagan 1967; Roger, Muñoz-Gómez, and Kamikawa 2017; Keeling 2010). This evolution has involved the dramatic reduction of the organelle genomes (mtDNA and ptDNA respectively; generally, organelle DNA or oDNA) (Roger, Muñoz-Gómez, and Kamikawa 2017; Keeling 2010; Hjort et al. 2010). This reduction has proceeded to different extents in different taxa (Giannakis et al. 2022; Johnston and Williams 2016; Janouškovec et al. 2017). Some jakobid protists retain over 60 protein-coding genes in their mtDNA (Lang et al. 1997) while some parasitic organisms retain only 3 and others have completely lost mtDNA (Hjort et al. 2010; Maciszewski and Karnkowska 2019; de Paula, Allen, and van der Giezen 2012). Some Rhodophyta contain around 200 protein-coding ptDNA genes (Janouškovec et al. 2013) while some parasitic plants contain only dozens (Naumann et al. 2016) and some reduced forms of plastids – like the apicoplasts found in Apicomplexans (McFadden 2011) – even fewer.

The diversity of organelle DNA gene profiles across eukaryotes raises, among others, two dual questions. First, what are the properties of a *gene* that make it more or less likely to be retained in a given species? And second, what are the properties of a *species* that make it more or less likely to retain a given gene? Numerous hypotheses for the first question have been proposed. It has been suggested that gene encoding hydrophobic products are preferentially retained in oDNA due to the difficulty of importing remotely-encoded hydrophobic proteins into the organelle (von Heijne 1986; Björkholm et al. 2015), including via mistargeting of mitochondrial products expressed in the cytosol – demonstrated across different genes (Björkholm et al. 2017). The colocation for redox regulation (CoRR) hypothesis proposes that retaining essential genes in organelles – where they can be expressed in response to local stimuli – supports local control of redox poise and organelle

behaviour rather than relying on slow and indirect nuclear transport (Allen 2015; Allen and Martin 2016; de Paula, Allen, and van der Giezen 2012). Economic considerations of the energy required by different encoding locations also support organelle retention of some genes (Kelly 2021). A role for organelle genes as long-term redox sensors, identifying bioenergetically competent cells, has been proposed (Wright, Murphy, and Turnbull 2009). Data-driven comparison of hypotheses using cross-eukaryotic genome data has provided support for the retention of genes encoding hydrophobic products, and products which are physically central in their functional complexes (supporting CoRR-like control) (Giannakis et al. 2022; Johnston and Williams 2016). Pronounced convergence in oDNA ribosomal protein-coding gene content across species has been found in support of CoRR-like pressures on gene retention (Maier et al. 2013).

Some of these mechanisms – likely in combination (Giannakis et al. 2022; Johnston and Williams 2016) – may explain some of the gene-to-gene variability in retention. But what of the dual species-to-species variability in number of oDNA genes retained? It is well known that many parasites atrophy organelle genomes to very reduced or even absent oDNA, likely due to a reduced energetic demand resulting from their exploitation of host metabolism (Keeling 2010; Hjort et al. 2010; Roger, Muñoz-Gómez, and Kamikawa 2017; Giannakis et al. 2022). Self-pollinating and clonal plant species typically transfer more oDNA genes to the nucleus; theoretical work has demonstrated that self-pollination accelerates such transfer when nuclear encoding is beneficial (Brandvain, Barker, and Wade 2007; Brandvain and Wade 2009). The coupled effects of nuclear genome reduction have been hypothesized as a reason for relative gene richness of red algal ptDNA (Qiu et al. 2017). The “limited transfer window” hypothesis places bounds on the time period during which a given taxa can transfer organelle genes (Barbrook, Howe, and Purton 2006).

Outside these specific cases, a general picture (if even applicable) has yet to fully emerge. One candidate is the “mutational hazard hypothesis” (Lynch, Koskella, and Schaack 2006), which ties gene retention to taxon-specific oDNA mutation rates, suggesting that organisms with low rates of oDNA mutation (for example, plants) can support more organelle-encoded content without risking deleterious effects of this content becoming mutated. This picture has support from some observations and competition from others (Smith 2016a), with several instances where the hypothesized relationship between mutation rate and coding content runs in a different way to that predicted. Another candidate contribution to a cross-species picture is an interpretation of CoRR which attempts to bridge intercellular and organismal scales. Here, when organelles are required to rapidly adapt to changing environmental conditions experienced by their “host”, they are predicted to retain more oDNA genes, to allow individual local control of organelles (Johnston 2019). Hence, organisms experiencing highly variable bioenergetic and metabolic demands from their environment would retain more oDNA genes, and those experiencing lower and more stable demands would retain fewer. This picture is supported to some extent by mathematical modelling (García-Pascual, Nordbotten, and Johnston 2022) and comparative analysis of links between ecological traits and oDNA gene counts (Giannakis, Richards, and Johnston 2023) – although we must point out that this picture does not and probably cannot explain the full diversity of oDNA retention profiles across eukaryotes. Although this theoretical picture predicts that similar factors shape – to some extent – mtDNA and ptDNA evolution, this prediction has to our knowledge yet to be tested with data. Here, while accepting that seeking too tight links between mitochondrial and plastid evolution can be misleading (Smith and Keeling 2015), we set out to explore whether a quantifiable relationship exists between mtDNA and ptDNA gene counts across eukaryotes.

Methods

Data gathering and curation

We used the pipeline from (Giannakis et al. 2022) to collect cross-eukaryote data on protein-coding oDNA genes from NCBI's Organelle Genome database (O'Leary et al. 2016) and to curate and unify annotations. This pipeline involves resolving the inconsistent annotation of genes across eukaryotic records via two branches. First, an unsupervised all-against-all BLAST comparison (Camacho et al. 2009) of all oDNA CDS (coding sequence) records is performed, followed by iterative hierarchical clustering and relabelling based on the BLAST outputs, to give sets of genes that form a connected network in the space of BLAST comparison metrics (regardless of original annotation). Second, a supervised process of relabelling annotations based on a manually compiled dictionary of annotation protocols across taxa. The relabelled annotations from these two approaches are then compared and any inconsistencies manually resolved.

NCBI's Common Taxonomy Tool (Federhen 2012) was used to estimate phylogenetic topology. The Encyclopedia of Life (Parr et al. 2014), Wikipedia (Wikipedia 2023), World Register of Marine Species (Ahyong et al. 2023), the USDA PLANTS Database (USDA and NRCS 2023), World Flora Online (WFO 2023), and AlgaeBase (Guiry and Guiry 2023) were used to collate ecological information on species.

Statistical analysis and visualisation

We used *phytools* (Revell 2012), *ape* (Paradis and Schliep 2019), and *phangorn* (Schliep 2011) for phylogenetic construction and analysis. For phylogenetic linear models (PLM) we used the *phylolm* package (Tung Ho and Ané 2014) for linear regression accounting for phylogenetic correlation with generalized estimating equations (Paradis and Claude 2002). We used *nlme* (DebRoy 2006) and *lme4* (Bates et al. 2015) for linear mixed models and *pheatmap* (Kolde 2019) for clustering and analysis. We used *Oncotree* (Szabo and Boucher 2008) for inference of progression. *igraph* (Csardi and Nepusz 2006) provided network analysis and visualization support. For visualization we used *ggplot2* (Wickham 2011), *ggpubr* (Kassambara 2020), *ggrepel* (Slowikowski 2021), *ggraph* (Pedersen 2020), *ggtree* (Yu et al. 2017), *ggtreeExtra* (Xu et al. 2021), and *ggVennDiagram* (Gao, Yu, and Cai 2021).

Results

Profiles of mitochondrial and chloroplast gene count across eukaryotes

We used the pipeline from (Giannakis et al. 2022) to collect cross-eukaryote data on protein-coding oDNA genes from NCBI's Organelle Genome database (O'Leary et al. 2016) and curate and unify annotations (see Methods). Following this pipeline, we found 205 unique species for which mitochondrial and plastid gene counts were available. 147 were Viridiplantae (green plants and algae), 30 Rhodophyta (red algae), 11 Bacillariophyta (diatoms), 9 Phaeophyceae (brown algae), 2 Apicomplexa (typically parasitic protists), and representatives of other algal groups: 2 Cryptophyceae, and 1 each of Bolidophyceae, Eustigmatophyceae, Glaucocystophyceae, and Haptista. A simple plot of mitochondrial (MT) gene count vs plastid (PT) gene count (Fig. 1A) shows considerable range and structure in the protein-coding gene count profiles. The Viridiplantae samples occupy a wide spread of MT counts at intermediate PT counts; Rhodophyta cluster at intermediate MT and very high PT counts; other algae occupy a high MT/high PT region.

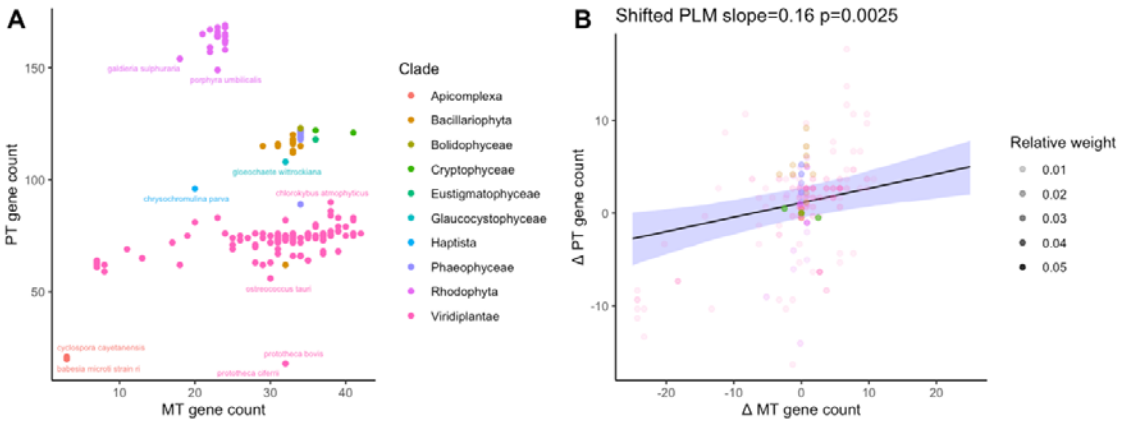


Figure 1. **MT and PT gene counts across eukaryotes.** (A) Raw protein-coding oDNA gene counts across species in our dataset, labelled by basal eukaryotic clade. Some species are labelled for illustration. (B) oDNA differences from clade average (ΔMT and ΔPT). Opacity of points in (B) is an illustration of their weighting in phylogenetic linear modelling (PLM): the inverse of the sum of their associated row in the variance-covariance matrix, reflecting a species' relatedness to other members of the dataset. Sets of closely related species provide less independent information and so have a lower weighting. A PLM fit suggests a significant (p = 0.0025) relationship between ΔMT and ΔPT, also found in a linear mixed model for raw oDNA counts (Supplementary Fig. 1).

Several notable outliers were found from this data summary – all with lower-than-expected PT gene counts. These were *Laminaria digitata* (sublittoral brown alga), *Nitzschia alba* (coldwater diatom), and *Prototheca* spp. (parasitic algae). *N. alba* and *Prototheca* are both anomalous in the sense that they are apochlorotic members of their respective clades – having abandoned photosynthesis, their plastid genomes have become reduced. *L. digitata* does not have this feature. Probing this case further, we found that although its chloroplast genome is described as containing 139 coding sequences (CDSs) (Rana et al. 2019), its NCBI sequence NC_044689.1 only reports 107 CDSs. As its closest relatives in our dataset all have around 140 cpDNA CDSs, we marked this sequence as reflecting a possible missing data artefact, where not all genes are present in the reference sequence. The following analysis removes these outlier sequences, remembering that apochlorotic species depart from the trends we will report.

Correlated departures from clade-wide gene count averages

To account for phylogenetic signal and pseudoreplication by relatedness (Paradis 2014; Maddison and FitzJohn 2015; Freckleton, Harvey, and Pagel 2002) in these observations, we changed variables to consider each species' difference in oDNA gene count from its clade's average, and performed phylogenetic linear modelling (PLM) accounting for relatedness. Thus, the mean MT and PT gene counts across a clade were computed, then subtracted from each species' individual counts to obtain a difference (ΔMT and ΔPT respectively). A relationship between ΔMT and ΔPT, robust to phylogenetic correction, would suggest that those species with lower-than-expected MT gene counts also have lower-than-expected PT gene counts, and vice versa. In turn, this could suggest that similar pressures shape MT and PT gene counts at the species level.

A reasonably clear positive relationship (slope=0.16, p=0.0025) exists between ΔMT and ΔPT (Fig. 1B), although it must immediately be noted that a relatively small amount of the variance in one organelle gene count is explained by the other (R^2 = 0.28). Supplementary Fig. 1 shows that, as expected, a linear mixed model with random effects assigned to clade shows a similar relationship.

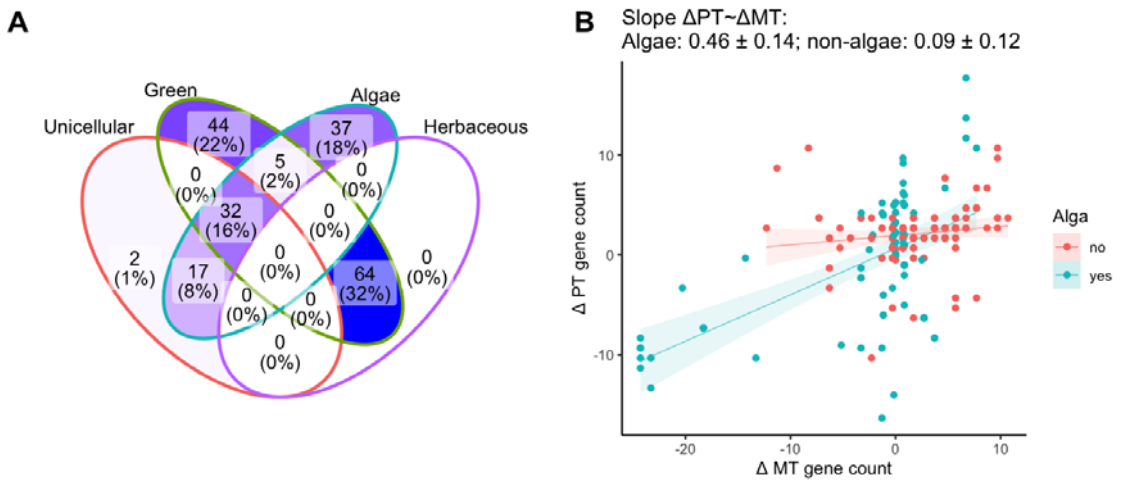


Figure 2. **Ecological influences in ΔMT-ΔPT relationship.** (A) Summary of the different classes of organism in the dataset. The two unicellular species that occupy no other sets are Apicomplexans. The herbaceous category is not applicable to algae or unicellular organisms. (B) ΔMT-ΔPT for organisms labelled as algae and organisms not labelled as algae. Statistically robust differences in MT-PT behaviour were not found for other arrangements of the features in (A).

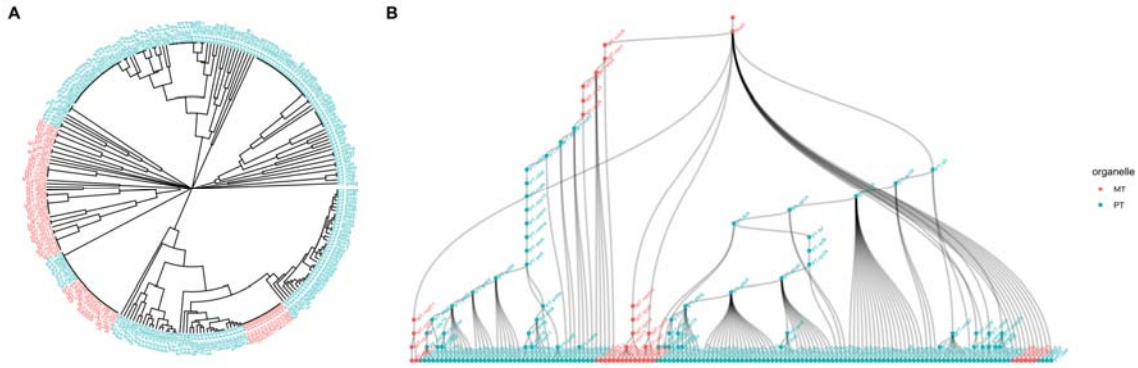


Figure 3. **Limited evidence for gene-gene relationships across organelle types.** (A) Hierarchical clustering of gene retention profiles in our dataset (Supplementary Fig. 2) produces a tree of relationships between genes that respects MT and PT differences. (B) Oncotree analysis of dependencies in MT-PT evolution, reporting inferred gene loss ordering in the dataset, shows mainly relationships within an organelle type (higher nodes are lost before connected lower nodes). Some MT genes (top, left of centre) precede a collection of PT genes, but the majority of genes cluster exclusively by organelle type.

To guard against artefacts from our chosen analysis method, we also analysed the raw MT and PT gene counts using PLM (without shifting by clade mean), and the relationship between ΔMT and ΔPT using a linear model (without phylogenetic correction via PLM). All variants of this process showed a positive relationship between mtDNA and ptDNA gene counts: slope 0.56, $p=4.9\times10^{-5}$ for PLM without clade mean shifting (equivalent to analysing the points in Fig. 1A accounting for relatedness); and slope 0.38, $p=4.2\times10^{-15}$ for a linear model after clade mean shifting (equivalent to equally weighting each point in Fig. 1B).

As Viridiplantae are the most sampled clade in our dataset, we also asked whether an MT-PT relationship was detectable within and without this clade. We used PLM to investigate the relationship between raw mtDNA and ptDNA gene counts in each case, finding significant

relationships both within Viridiplantae (slope=0.15, p=0.0059) and without Viridiplantae (slope=2.0, $p=6.5\times10^{-6}$).

Ecological and species-specific factors

Following work proposing ecological factors as shapers of oDNA evolution (García-Pascual, Nordbotten, and Johnston 2022; Giannakis, Richards, and Johnston 2023), we asked whether ecological factors further separated substructure in the MT-PT space (Fig. 2). Following findings in (Giannakis, Richards, and Johnston 2023), aligned with the features that most broadly distinguished entries in our dataset, we tested for differences between organisms with different characteristics: algae vs non-algae, unicellular vs multicellular organisms, green plants/algae vs others, herbaceous plants vs others, annual vs perennial lifestyle. For clade-corrected oDNA counts (ΔMT and ΔPT), a linear mixed model with random and gradients assigned to the algal category was slightly favoured by AIC over the fixed linear model (1154.7 vs 1157.5 for the fixed model). In other words, there is some statistical support for a picture where the relationship between ΔMT and ΔPT differs between algae and non-algae (Fig. 2B). We found no statistically robust differences in oDNA behaviour for other categorical differences between organisms.

Gene-level relationships in mtDNA and ptDNA evolution

We next asked whether there was evidence for evolutionary relationships between specific MT and PT genes – for example, is there a particular PT gene that is always lost when a particular MT gene is also lost? To this end we used hierarchical clustering of the dataset to explore interdependence of PT-MT gene patterns (Supplementary Fig. 2). We also used Oncotree, an algorithm for inferring relationships in progressive evolutionary processes (Szabo and Boucher 2008), to explore evidence for gene-specific dependencies in our data. Oncotree is designed for and applied to features in cancer progression but fundamentally deals with progressive evolutionary changes in binary traits.

We found no evidence of gene-gene correlations or conditional dependencies via hierarchical clustering of the data (Fig. 3A), with MT and PT genes clustering separately across the retention profiles in our dataset, rather than clusters of MT and PT genes together which could indicate subset-specific correlations. Oncotree suggested some consistent ordering behaviour across organelles, where the loss of some MT genes (*ccmb, rps1, atp4, atp1, rps3, atp9*) is inferred to characteristically precede the loss of a collection of PT genes (Fig. 3B). This MT subset are not consistently among the most or least retained of organelle genes – their “retention indices” from (Giannakis et al. 2022) recording how many other genes are typically lost before them, take a range of intermediate values. The “dependent” PT subset overwhelmingly comprises highly-retained genes (Supp. Fig. 3). Most other genes from both organelles clustered separately in this analysis too, suggesting only limited gene-specific relationships between organelles.

Discussion

Our work suggests that, controlling both for phylogenetic signal and the relatedness of individuals, there is some support for a picture where MT and PT gene counts are connected. This does not appear to be manifest through pronounced cross-organelle interactions between specific genes (where, for example, the loss of MT gene X always results in the loss of PT gene Y). Rather, it appears that the drivers of reductive evolution apply comparably, but largely independently, to the two organelles. These drivers appear to be connected to some aspects of organismal lifestyle – algae and non-algae, for example, display different MT-PT relationships.

This picture is compatible with recent results on both gene-specific (Giannakis et al. 2022; Johnston and Williams 2016) and species-specific (García-Pascual, Nordbotten, and Johnston 2022) drivers of oDNA genome evolution. Here, the same combination of features – hydrophobicity, GC content, and the centrality of a protein product in its functional complex

– predict a “retention index” describing how readily a gene is lost in either mtDNA or ptDNA. In parallel, the demands placed on organelle control by a species’ environment determine how many oDNA genes it retains. Broadly, those with higher retention indices are retained even by species in less challenging environments; species in more challenging environments retain genes with progressively lower retention indices. There are, of course, many more determinants of the oDNA retention profile of individual species – from the outcomes of stochastic evolutionary processes to population genetic considerations, and doubtless other mechanisms which the above picture does not consider. Correspondingly, the gene-specific models in (Giannakis, Richards, and Johnston 2023; Johnston and Williams 2016) can explain just over half of the variance in gene retention indices (R^2 = 0.51-0.60), and the relationship between MT and PT found here is certainly not clear-cut, with just over a quarter of variance in ΔPT explained by ΔMT (R^2 = 0.28; Fig. 1B). This figure is of course after removing the apochlorotic and potentially misannotated outliers in our data, the inclusion of which would lead to still less variance being explained.

In correcting for systematic differences in gene count between deep-branching clades (Giannakis et al. 2022; Janouškovec et al. 2017) and additionally for the phylogenetic relationship between samples, we have attempted to align with the picture in (Maddison and FitzJohn 2015), of both identifying independent subsets of data and then accounting “internally” for phylogenetic signal. However, how best to account for dependencies across phylogenetically-embedded data is a topic of substantial discussion (Uyeda, Zenil-Ferguson, and Pennell 2018; Losos 2011); even well-principled comparative methods have shortcomings in the face of singular evolutionary events (Uyeda, Zenil-Ferguson, and Pennell 2018) and imperfect phylogenetic estimates (as we have here) and the distinction between pattern and process can challenge the application of these methods (Losos 2011) – discussed further in an oDNA context in (Giannakis, Richards, and Johnston 2023). In taking different combinations of clade-based and phylogenetic correction, and considering different subtrees of our data, we have attempted to be robust to particular choices of method. However, as more organelle genome data becomes available, testing alternative pictures like conditional dependencies between variables (Uyeda, Zenil-Ferguson, and Pennell 2018) will be an interesting line of inquiry.

The timescale of genome reduction must also be discussed. (Janouškovec et al. 2017), for example, give a compelling picture where the majority of mtDNA gene loss occurred early in the history of most eukaryotic clades, with modern-day mtDNA being more (in some cases almost completely) static. Despite evidence of ongoing movement of genetic information from oDNA to the nucleus (Timmis et al. 2004) – for example, nuclear mitochondrial sequences (NUMTs) from mtDNA in humans (Wei et al. 2022) and very frequent ptDNA transfer in plants (Stegemann et al. 2003; Bock and Timmis 2008) – the gene profiles of many eukaryotic clades seems quite fixed. The two pictures can be reconciled by noting that although NUMTs and NUPTs can enter the nuclear genome, the functionalization of these fragments into a working gene, and the subsequent removal of the organelle-encoded precursor, presumably involve many more genetic and regulatory steps and can reasonably be considered a rare event in many species – as in the limited transfer window hypothesis mentioned previously (Barbrook, Howe, and Purton 2006). Overall, the patterns in gene content observed across modern oDNA samples were presumably largely established in the rather distant evolutionary past. Hence, the processes we study here are historical; and while modern-day oDNA profiles are clearly compatible with modern-day environments, it was not modern-day environments that originally shaped them.

We have only considered protein-coding gene counts across species – a very coarse-grained structural feature of oDNA. Tremendous diversity also exists in oDNA length and proportion of coding content, genetic structure (gene orderings and repeat regions), and physical structure (linear, branching, circular, and fragmented molecules) of oDNA across eukaryotes (Roger, Muñoz-Gómez, and Kamikawa 2017; Keeling 2010). Explanations for

this diversity include the mutational hazard hypothesis (Lynch, Koskella, and Schaack 2006; Smith 2016a), effects of oDNA recombination in some lineages (Edwards et al. 2021), and recent work suggesting a functional link between gene order and the control of gene expression in metazoa (Shtolz and Mishmar 2023); the interplay (or independence) of these mechanisms with the environmental influences we propose is an interesting target for future theoretical exploration.

Despite advances in sequencing protist oDNA, many clades are unrepresented in our dataset or only represented by a single species. Complete organelle records exist in NCBI for over 9k ptDNA sequences and over 2k mtDNA sequences – but the large majorities of these are from green plants and animals respectively (Smith 2016b). Organelle genome information from less represented clades – especially, for this study, photosynthetic protists – would be immensely valuable in verifying and refining these theories of organelle co-evolution. At the same time, such species play essential roles in marine and other ecosystems, and their photosynthetic and metabolic capacity and responses to environmental change are of profound importance on the warming and increasingly dynamic planet (Beardall and Raven 2004). More generally, oDNA features play an important role in the adaptability and evolvability of organisms in changing environments (Radzvilavicius and Johnston 2022; 2020), and environmental influences on photosynthesis and respiration are central to the biosphere's response to climate change (Dusenge, Duarte, and Way 2019). Understanding these essential genomes further, and the advantages and disadvantages of gene retention in different environments, is highly desirable in the study of diverse eukaryotic responses to a changing world.

Data and code availability

All data and code used in analysis and visualization is available at github.com/StochasticBiology/odna-coevolution/

Acknowledgements

LR was supported by the BBSRC via the MIBTP Doctoral Training Scheme. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 805046 (EvoConBiO) to IGJ). KD was supported by the Trond Mohn Foundation [project HyperEvol under grant agreement No. TMS2021TMT09], through the Centre for Antimicrobial Resistance in Western Norway (CAMRIA) [TMS2020TMT11].

References

Ahyong, S., C.B. Boyko, N. Bailly, J. Bernot, R. Bieler, S.N. Brandão, M. Daly, et al. 2023. 'World Register of Marine Species (WoRMS)'. WoRMS Editorial Board. https://www.marinespecies.org.

Allen, John F. 2015. 'Why Chloroplasts and Mitochondria Retain Their Own Genomes and Genetic Systems: Colocation for Redox Regulation of Gene Expression'. *Proceedings of the National Academy of Sciences* 112 (33): 10231–38. https://doi.org/10.1073/pnas.1500012112.

Allen, John F., and William F. Martin. 2016. 'Why Have Organelles Retained Genomes?' *Cell Systems* 2 (2): 70–72. https://doi.org/10.1016/j.cels.2016.02.007.

Barbrook, Adrian C., Christopher J. Howe, and Saul Purton. 2006. 'Why Are Plastid Genomes Retained in Non-Photosynthetic Organisms?' *Trends in Plant Science* 11 (2): 101–8. https://doi.org/10.1016/j.tplants.2005.12.004.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. 'Fitting Linear Mixed-Effects Models Using Lme4'. *Journal of Statistical Software* 67 (October): 1–48. https://doi.org/10.18637/jss.v067.i01.

Beardall, John, and John A. Raven. 2004. 'The Potential Effects of Global Climate Change on Microalgal Photosynthesis, Growth and Ecology'. *Phycologia* 43 (1): 26–40. https://doi.org/10.2216/i0031-8884-43-1-26.1.

Björkholm, Patrik, Andreas M. Ernst, Erik Hagström, and Siv G. E. Andersson. 2017. 'Why Mitochondria Need a Genome Revisited'. *FEBS Letters* 591 (1): 65–75. https://doi.org/10.1002/1873-3468.12510.

Björkholm, Patrik, Ajith Harish, Erik Hagström, Andreas M. Ernst, and Siv GE Andersson. 2015. 'Mitochondrial Genomes Are Retained by Selective Constraints on Protein Targeting'. *Proceedings of the National Academy of Sciences* 112 (33): 10154–61.

Bock, Ralph, and Jeremy N. Timmis. 2008. 'Reconstructing Evolution: Gene Transfer from Plastids to the Nucleus'. *BioEssays* 30 (6): 556–66. https://doi.org/10.1002/bies.20761.

Brandvain, Yaniv, Michael S. Barker, and Michael J. Wade. 2007. 'Gene Co-Inheritance and Gene Transfer'. *Science (New York, N.Y.)* 315 (5819): 1685. https://doi.org/10.1126/science.1134789.

Brandvain, Yaniv, and Michael J. Wade. 2009. 'The Functional Transfer of Genes From the Mitochondria to the Nucleus: The Effects of Selection, Mutation, Population Size and Rate of Self-Fertilization'. *Genetics* 182 (4): 1129–39. https://doi.org/10.1534/genetics.108.100024.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. 'BLAST+: Architecture and Applications'. *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Csardi, Gabor, and Tamas Nepusz. 2006. 'The Igraph Software Package for Complex Network Research'. *InterJournal, Complex Systems* 1695 (5): 1–9.

DebRoy, S. 2006. 'Nlme: Linear and Nonlinear Mixed Effects Models'. *R Package*.

Dusenge, Mirindi Eric, André Galvao Duarte, and Danielle A. Way. 2019. 'Plant Carbon Metabolism and Climate Change: Elevated CO2 and Temperature Impacts on Photosynthesis, Photorespiration and Respiration'. *New Phytologist* 221 (1): 32–49. https://doi.org/10.1111/nph.15283.

Edwards, David M., Ellen C. Røyrvik, Joanna M. Chustecki, Konstantinos Giannakis, Robert C. Glastad, Arunas L. Radzvilavicius, and Iain G. Johnston. 2021. 'Avoiding Organelle Mutational Meltdown across Eukaryotes with or without a Germline Bottleneck'. *PLOS Biology* 19 (4): e3001153. https://doi.org/10.1371/journal.pbio.3001153.

Federhen, Scott. 2012. 'The NCBI Taxonomy Database'. *Nucleic Acids Research* 40 (D1): D136–43. https://doi.org/10.1093/nar/gkr1178.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. 'Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence.' *The American Naturalist* 160 (6): 712–26. https://doi.org/10.1086/343873.

Gao, Chun-Hui, Guangchuang Yu, and Peng Cai. 2021. 'ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram'. *Frontiers in Genetics*, 1598.

García-Pascual, Belén, Jan M. Nordbotten, and Iain G. Johnston. 2022. 'Cellular and Environmental Dynamics Influence Species-Specific Extents of Organelle Gene Retention'. bioRxiv. https://doi.org/10.1101/2022.10.17.512581.

Giannakis, Konstantinos, Samuel J. Arrowsmith, Luke Richards, Sara Gasparini, Joanna M. Chustecki, Ellen C. Røyrvik, and Iain G. Johnston. 2022. 'Evolutionary Inference across Eukaryotes Identifies Universal Features Shaping Organelle Gene Retention'. *Cell Systems* 13 (11): 874-884.e5. https://doi.org/10.1016/j.cels.2022.08.007.

Giannakis, Konstantinos, Luke Richards, and Iain Johnston. 2023. 'Ecological Predictors of Organelle Genome Evolution: Phylogenetic Correlations with Taxonomically Broad, Sparse, Unsystematized Data'. bioRxiv. https://doi.org/10.1101/2023.08.11.553003.

Guiry, M. D., and G. M. Guiry. 2023. 'AlgaeBase. World-Wide Electronic Publication, National University of Ireland, Galway.' 2023. https://www.algaebase.org.

Heijne, Gunnar von. 1986. 'Why Mitochondria Need a Genome'. *FEBS Letters* 198 (1): 1–4.

Hjort, Karin, Alina V. Goldberg, Anastasios D. Tsaousis, Robert P. Hirt, and T. Martin Embley. 2010. 'Diversity and Reductive Evolution of Mitochondria among Microbial Eukaryotes'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1541): 713–27. https://doi.org/10.1098/rstb.2009.0224.

Janouškovec, Jan, Shao-Lun Liu, Patrick T. Martone, Wilfrid Carré, Catherine Leblanc, Jonas Collén, and Patrick J. Keeling. 2013. 'Evolution of Red Algal Plastid Genomes: Ancient Architectures, Introns, Horizontal Gene Transfer, and Taxonomic Utility of Plastid Markers'. *PLOS ONE* 8 (3): e59001. https://doi.org/10.1371/journal.pone.0059001.

Janouškovec, Jan, Denis V. Tikhonenkov, Fabien Burki, Alexis T. Howe, Forest L. Rohwer, Alexander P. Mylnikov, and Patrick J. Keeling. 2017. 'A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction'. *Current Biology* 27 (23): 3717-3724.e5. https://doi.org/10.1016/j.cub.2017.10.051.

Johnston, Iain G. 2019. 'Tension and Resolution: Dynamic, Evolving Populations of Organelle Genomes within Plant Cells'. *Molecular Plant* 12 (6): 764–83. https://doi.org/10.1016/j.molp.2018.11.002.

Johnston, Iain G., and Ben P. Williams. 2016. 'Evolutionary Inference across Eukaryotes Identifies Specific Pressures Favoring Mitochondrial Gene Retention'. *Cell Systems* 2 (2): 101–11. https://doi.org/10.1016/j.cels.2016.01.013.

Kassambara, Alboukadel. 2020. 'Ggpubr:"Ggplot2" Based Publication Ready Plots'. *R Package Version 0.4. 0* 438.

Keeling, Patrick J. 2010. 'The Endosymbiotic Origin, Diversification and Fate of Plastids'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1541): 729–48. https://doi.org/10.1098/rstb.2009.0103.

Kelly, Steven. 2021. 'The Economics of Organellar Gene Loss and Endosymbiotic Gene Transfer'. *Genome Biology* 22 (1): 345. https://doi.org/10.1186/s13059-021-02567-w.

Kolde, Raivo. 2019. 'Pheatmap: Pretty Heatmaps. R Package Version 1.0. 12'.

Lang, B. Franz, Gertraud Burger, Charles J. O'Kelly, Robert Cedergren, G. Brian Golding, Claude Lemieux, David Sankoff, Monique Turmel, and Michael W. Gray. 1997. 'An Ancestral Mitochondrial DNA Resembling a Eubacterial Genome in Miniature'. *Nature* 387 (6632): 493–97.

Losos, Jonathan B. 2011. 'Seeing the Forest for the Trees: The Limitations of Phylogenies in Comparative Biology.' *The American Naturalist* 177 (6): 709–27. https://doi.org/10.1086/660020.

Lynch, Michael, Britt Koskella, and Sarah Schaack. 2006. 'Mutation Pressure and the Evolution of Organelle Genomic Architecture'. *Science* 311 (5768): 1727–30. https://doi.org/10.1126/science.1118884.

Maciszewski, Kacper, and Anna Karnkowska. 2019. 'Should I Stay or Should I Go? Retention and Loss of Components in Vestigial Endosymbiotic Organelles'. *Current Opinion in Genetics & Development*, Evolutionary genetics, 58–59 (October): 33–39. https://doi.org/10.1016/j.gde.2019.07.013.

Maddison, Wayne P., and Richard G. FitzJohn. 2015. 'The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters'. *Systematic Biology* 64 (1): 127–36. https://doi.org/10.1093/sysbio/syu070.

Maier, Uwe-G, Stefan Zauner, Christian Woehle, Kathrin Bolte, Franziska Hempel, John F. Allen, and William F. Martin. 2013. 'Massively Convergent Evolution for Ribosomal Protein Gene Content in Plastid and Mitochondrial Genomes'. *Genome Biology and Evolution* 5 (12): 2318–29. https://doi.org/10.1093/gbe/evt181.

McFadden, Geoffrey Ian. 2011. 'The Apicoplast'. *Protoplasma* 248 (4): 641–50. https://doi.org/10.1007/s00709-010-0250-5.

Naumann, Julia, Joshua P. Der, Eric K. Wafula, Samuel S. Jones, Sarah T. Wagner, Loren A. Honaas, Paula E. Ralph, et al. 2016. 'Detecting and Characterizing the Highly Divergent Plastid Genome of the Nonphotosynthetic Parasitic Plant Hydnora Visseri (Hydnoraceae)'. *Genome Biology and Evolution* 8 (2): 345–63. https://doi.org/10.1093/gbe/evv256.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. 'Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation'. *Nucleic Acids Research* 44 (D1): D733–45. https://doi.org/10.1093/nar/gkv1189.

Paradis, Emmanuel. 2014. 'An Introduction to the Phylogenetic Comparative Method'. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, 3–18.

Paradis, Emmanuel, and Julien Claude. 2002. 'Analysis of Comparative Data Using Generalized Estimating Equations'. *Journal of Theoretical Biology* 218 (2): 175–85. https://doi.org/10.1006/jtbi.2002.3066.

Paradis, Emmanuel, and Klaus Schliep. 2019. 'Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R'. *Bioinformatics* 35 (3): 526–28. https://doi.org/10.1093/bioinformatics/bty633.

Parr, Cynthia S., Nathan Wilson, Patrick Leary, Katja S. Schulz, Kristen Lans, Lisa Walley, Jennifer A. Hammock, et al. 2014. 'The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth'. *Biodiversity Data Journal*, no. 2 (April): e1079. https://doi.org/10.3897/BDJ.2.e1079.

Paula, Wilson B. M. de, John F. Allen, and Mark van der Giezen. 2012. 'Mitochondria, Hydrogenosomes and Mitosomes in Relation to the CoRR Hypothesis for Genome Function and Evolution'. In *Organelle Genetics: Evolution of Organelle Genomes and Gene Expression*, edited by Charles E. Bullerwell, 105–19. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-22380-8_5.

Pedersen, Thomas Lin. 2020. 'Ggraph: An Implementation of Grammar of Graphics for Graphs and Networks'. *R Package Version* 2 (3): 1.

Qiu, Huan, Jun Mo Lee, Hwan Su Yoon, and Debashish Bhattacharya. 2017. 'Hypothesis: Gene-Rich Plastid Genomes in Red Algae May Be an Outcome of Nuclear Genome Reduction'. *Journal of Phycology* 53 (3): 715–19. https://doi.org/10.1111/jpy.12514.

Radzvilavicius, Arunas L., and Iain G. Johnston. 2020. 'Paternal Leakage of Organelles Can Improve Adaptation to Changing Environments'. *bioRxiv*, 2020–12.

———. 2022. 'Organelle Bottlenecks Facilitate Evolvability by Traversing Heteroplasmic Fitness Valleys'. *Frontiers in Genetics* 13: 974472.

Rana, Shivani, Klaus Valentin, Inka Bartsch, and Gernot Glöckner. 2019. 'Loss of a Chloroplast Encoded Function Could Influence Species Range in Kelp'. *Ecology and Evolution* 9 (15): 8759–70. https://doi.org/10.1002/ece3.5428.

Revell, Liam J. 2012. 'Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things)'. *Methods in Ecology and Evolution*, no. 2: 217–23.

Roger, Andrew J., Sergio A. Muñoz-Gómez, and Ryoma Kamikawa. 2017. 'The Origin and Diversification of Mitochondria'. *Current Biology: CB* 27 (21): R1177–92. https://doi.org/10.1016/j.cub.2017.09.015.

Sagan, Lynn. 1967. 'On the Origin of Mitosing Cells'. *Journal of Theoretical Biology* 14 (3): 225-IN6. https://doi.org/10.1016/0022-5193(67)90079-3.

Schliep, Klaus Peter. 2011. 'Phangorn: Phylogenetic Analysis in R'. *Bioinformatics* 27 (4): 592–93. https://doi.org/10.1093/bioinformatics/btq706.

Shtolz, Noam, and Dan Mishmar. 2023. 'The Metazoan Landscape of Mitochondrial DNA Gene Order and Content Is Shaped by Selection and Affects Mitochondrial Transcription'. *Communications Biology* 6 (1): 1–15. https://doi.org/10.1038/s42003-023-04471-4.

Slowikowski, K. 2021. 'Ggrepel: Automatically Position Non-Overlapping Text Labels with'ggplot2'. R Package Version 0.9. 1, 2021'.

Smith, David Roy. 2016a. 'The Mutational Hazard Hypothesis of Organelle Genome Evolution: 10 Years On'. *Molecular Ecology* 25 (16): 3769–75. https://doi.org/10.1111/mec.13742.

———. 2016b. 'The Past, Present and Future of Mitochondrial Genomics: Have We Sequenced Enough mtDNAs?' *Briefings in Functional Genomics* 15 (1): 47–54. https://doi.org/10.1093/bfgp/elv027.

Smith, David Roy, and Patrick J. Keeling. 2015. 'Mitochondrial and Plastid Genome Architecture: Reoccurring Themes, but Significant Differences at the Extremes'. *Proceedings of the National Academy of Sciences* 112 (33): 10177–84. https://doi.org/10.1073/pnas.1422049112.

Stegemann, Sandra, Stefanie Hartmann, Stephanie Ruf, and Ralph Bock. 2003. 'High-Frequency Gene Transfer from the Chloroplast Genome to the Nucleus'. *Proceedings of the National Academy of Sciences* 100 (15): 8828–33. https://doi.org/10.1073/pnas.1430924100.

Szabo, Aniko, and Kenneth M. Boucher. 2008. 'Oncogenetic Trees'. In *Handbook of Cancer Models with Applications*, 1–24. World Scientific.

Timmis, Jeremy N., Michael A. Ayliffe, Chun Y. Huang, and William Martin. 2004. 'Endosymbiotic Gene Transfer: Organelle Genomes Forge Eukaryotic Chromosomes'. *Nature Reviews Genetics* 5 (2): 123–35. https://doi.org/10.1038/nrg1271.

Tung Ho, Lam si, and Cécile Ané. 2014. 'A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models'. *Systematic Biology* 63 (3): 397–408. https://doi.org/10.1093/sysbio/syu005.

USDA, and NRCS. 2023. 'USDA Plants Database, National Plant Data Team, Greensboro, NC USA.' 2023. https://plants.usda.gov/.

Uyeda, Josef C, Rosana Zenil-Ferguson, and Matthew W Pennell. 2018. 'Rethinking Phylogenetic Comparative Methods'. *Systematic Biology* 67 (6): 1091–1109. https://doi.org/10.1093/sysbio/syy031.

Wei, Wei, Katherine R. Schon, Greg Elgar, Andrea Orioli, Melanie Tanguy, Adam Giess, Marc Tischkowitz, Mark J. Caulfield, and Patrick F. Chinnery. 2022. 'Nuclear-Embedded Mitochondrial DNA Sequences in 66,083 Human Genomes'. *Nature* 611 (7934): 105–14.

WFO. 2023. 'World Flora Online'. 2023. http://www.worldfloraonline.org/.

Wickham, Hadley. 2011. 'Ggplot2'. *WIREs Computational Statistics* 3 (2): 180–85. https://doi.org/10.1002/wics.147.

Wikipedia. 2023. 'Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc.' 2023. https://www.wikipedia.org.

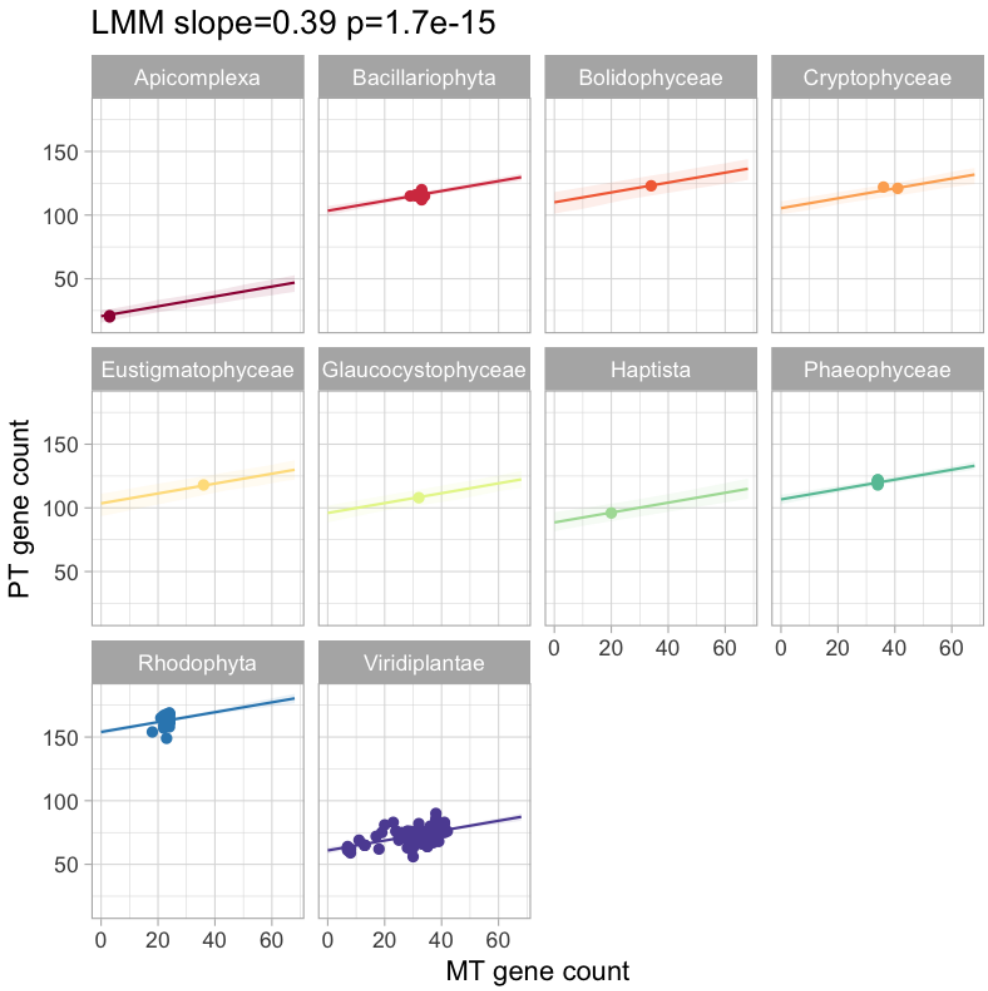
Wright, Alan F., Michael P. Murphy, and Douglass M. Turnbull. 2009. 'Do Organellar Genomes Function as Long-Term Redox Damage Sensors?' *Trends in Genetics* 25 (6): 253–61. https://doi.org/10.1016/j.tig.2009.04.006.

Xu, Shuangbin, Zehan Dai, Pingfan Guo, Xiaocong Fu, Shanshan Liu, Lang Zhou, Wenli Tang, et al. 2021. 'ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data'. *Molecular Biology and Evolution* 38 (9): 4039–42. https://doi.org/10.1093/molbev/msab166.

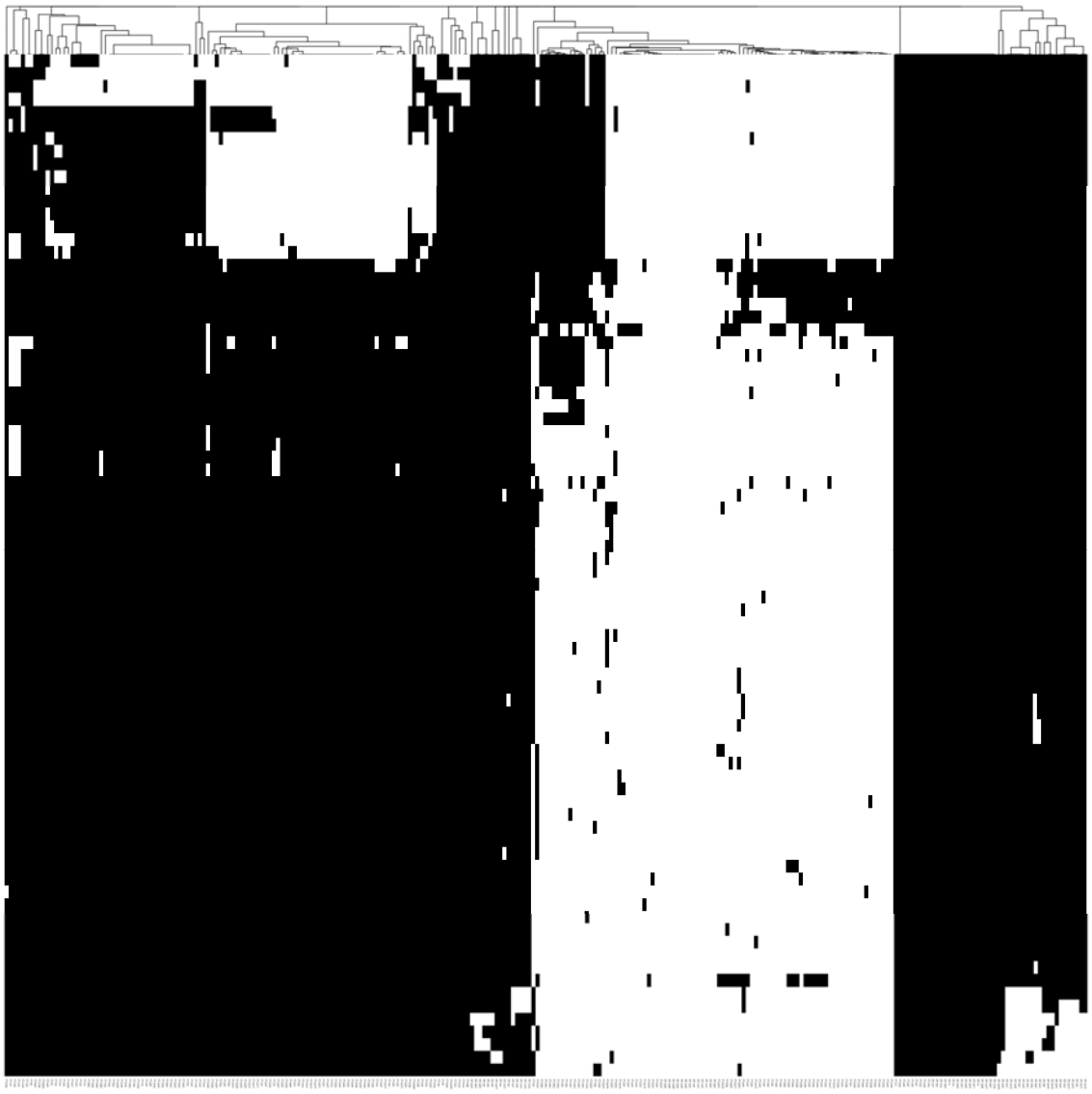
Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. 'Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data'. *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Zimorski, Verena, Chuan Ku, William F Martin, and Sven B Gould. 2014. 'Endosymbiotic Theory for Organelle Origins'. *Current Opinion in Microbiology*, Growth and development: eukaryotes/ prokaryotes, 22 (December): 38–48. https://doi.org/10.1016/j.mib.2014.09.008.

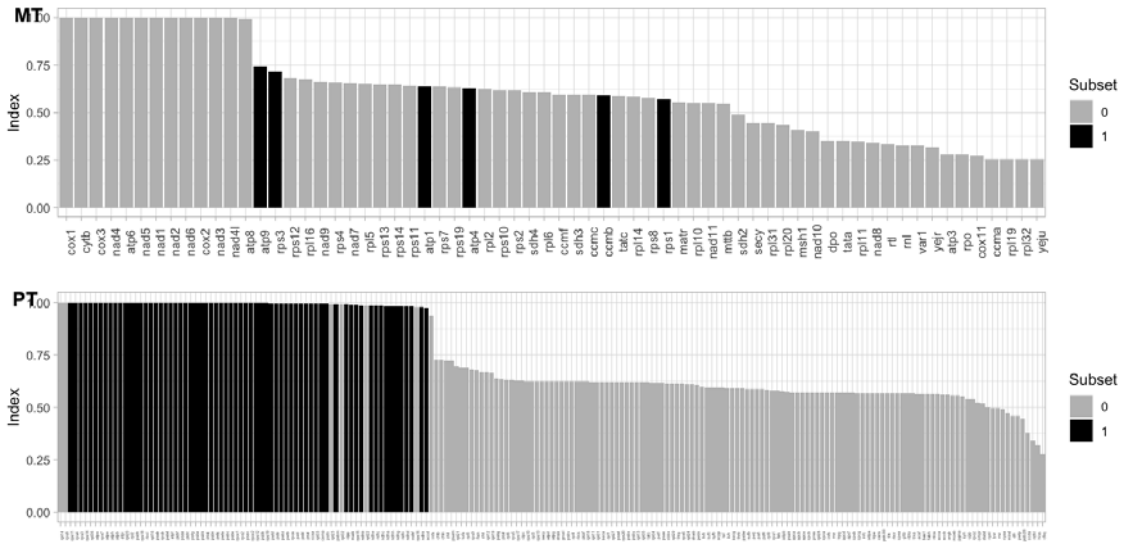
Supplementary Information



Supplementary Figure 1. Linear mixed model for MT-PT relationship in raw gene counts, with random effects assigned to eukaryotic clade.



Supplementary Figure 2. Heatmap from clustering in the combined MT-PT dataset; rows are retention profiles, columns are genes; black pixels are gene presence and white are gene absence. The dendrogram for gene clustering is reported in Fig. 3A.



Supplementary Figure 3. The subset of MT genes that Oncotree predicts to be lost before a subset of PT genes. Plots give the ranked "retention index" across oDNA genes from (Giannakis et al. 2022); genes with higher retention indices are inferred to be lost after genes with lower indices. The members of each subset identified by Oncotree are marked in black.