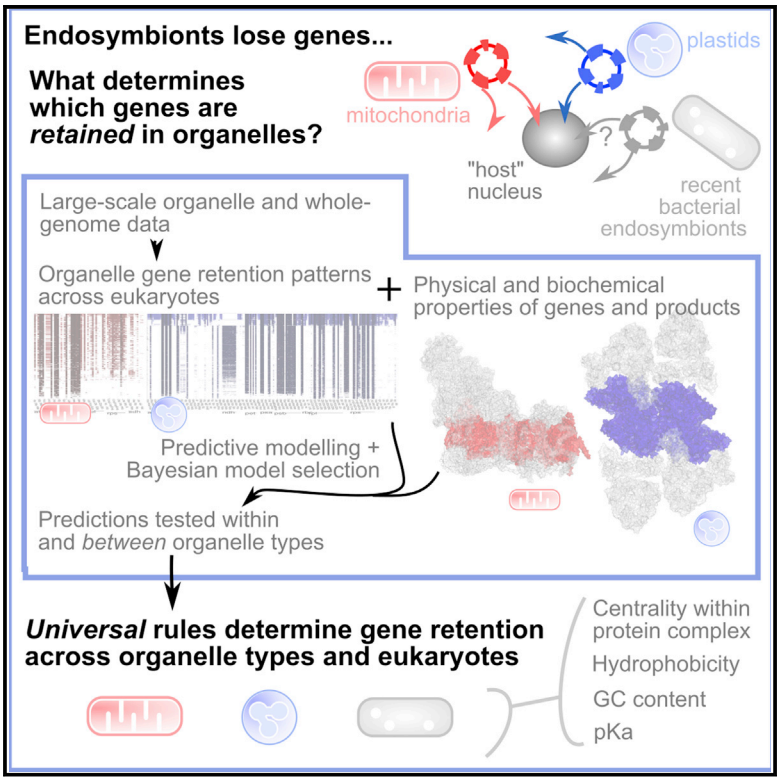
Cell Systems

Article

Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention

Graphical abstract



Highlights

- Predictors of organelle gene retention patterns quantified across eukaryotes
- Bayesian model selection harnesses genomic data to dissect likely mechanisms
- Gene retention in mtDNA, ptDNA, and recent endosymbionts follow the same rules
- Subunit binding energy and biochemistry predict retention, with hydrophobicity

Authors

Konstantinos Giannakis, Samuel J. Arrowsmith, Luke Richards, Sara Gasparini, Joanna M. Chustecki, Ellen C. Røyrvik, Iain G. Johnston

Correspondence

iain.johnston@uib.no

In brief

Why do mitochondria and plastids retain genes, and might the reasons be the same across all organelles? Giannakis and Arrowsmith et al. combine large-scale bioinformatic analysis with quantitative modelling of different possible mechanisms to find universal "rules" of how gene features govern the patterns of organelle genome evolution across eukaryotes.

Giannakis et al., 2022, Cell Systems *13*, 1–11
November 16, 2022
https://doi.org/10.1016/j.cels.2022.08.007

CellPress

Cell Systems

CellPress
OPEN ACCESS

Article

Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention

Konstantinos Giannakis,[1,8] Samuel J. Arrowsmith,[2,8] Luke Richards,[3] Sara Gasparini,[4] Joanna M. Chustecki,[5] Ellen C. Røyrvik,[6] and Iain G. Johnston[1,7,9,*]

[1]Department of Mathematics, University of Bergen, Bergen, Norway
[2]CNRS UMR7156, Génétique moléculaire, génomique, microbiologie (GMGM), Université de Strasbourg, Strasbourg, France
[3]Department of Life Sciences, University of Warwick, Coventry, UK
[4]Birkeland Centre for Space Science, Department of Physics and Technology, University of Bergen, Bergen, Norway
[5]School of Biosciences, University of Birmingham, Birmingham, UK
[6]Department of Clinical Sciences, University of Bergen, Bergen, Norway
[7]Computational Biology Unit, University of Bergen, Bergen, Norway
[8]These authors contributed equally
[9]Lead contact
*Correspondence: iain.johnston@uib.no
https://doi.org/10.1016/j.cels.2022.08.007

SUMMARY

Mitochondria and plastids power complex life. Why some genes and not others are retained in their organelle DNA (oDNA) genomes remains a debated question. Here, we attempt to identify the properties of genes and associated underlying mechanisms that determine oDNA retention. We harness over 15k oDNA sequences and over 300 whole genome sequences across eukaryotes with tools from structural biology, bioinformatics, machine learning, and Bayesian model selection. Previously hypothesized features, including the hydrophobicity of a protein product, and less well-known features, including binding energy centrality within a protein complex, predict oDNA retention across eukaryotes, with additional influences of nucleic acid and amino acid biochemistry. Notably, the same features predict retention in both organelles, and retention models learned from one organelle type quantitatively predict retention in the other, supporting the universality of these features—which also distinguish gene profiles in more recent, independent endosymbiotic relationships. A record of this paper's transparent peer review process is included in the supplemental information.

INTRODUCTION

Mitochondria and plastids (the broader class of organelles of which chloroplasts are one type) are bioenergetic organelles derived from the ancient endosymbiotic acquisition of bacterial precursors (Martin et al., 2015). The subsequent co-evolution of mitochondria and plastids with their host cells has shaped complex life (Lane and Martin, 2010; Hohmann-Marriott and Blankenship, 2011; Booth and Doolittle, 2015). Across eukaryotes, the genomes of the original endosymbionts (estimated to have contained thousands of genes; Boussau et al., 2004) have been dramatically reduced through evolutionary time (Blanchard and Lynch, 2000; Adams and Palmer, 2003; Martin et al., 2015). Genes have either been lost completely or transferred to the "host" cell nucleus, so that modern-day organelle DNA (oDNA) contains few genes. Many differences exist between the nuclear and organelle compartments as encoding environments (Bullerwell, 2011), and some of these can vary dramatically across eukaryotic taxa—including physical structure, gene density, presence of introns, and capacity for recombination (Edwards et al., 2021). Some more general differences between oDNA and nuclear DNA (nDNA) include cellular copy number (oDNA is often present in ploidy of hundreds or thousands), mutation rate (oDNA is often subject to a higher sequence mutation rate than nDNA, although not in plants; rates of structural change differ again; Johnston and Burgstaller, 2019), epigenetics (marking of oDNA, for example, appears much more limited than nDNA), and expression through different machinery and mechanisms (mitochondria and chloroplasts have specific, dedicated polymerases, ribosomes, and tRNAs). These differences mean that gene transfer from organelles has profound implications for the balance of control between the nucleus and endosymbiont and the inheritance and maintenance of vital genetic information (Sloan et al., 2018).

Selective pressures favoring organelle gene transfer are largely agreed upon (Adams and Palmer, 2003). Nuclear encoding allows recombination to avoid Muller's ratchet (the irreversible buildup of damaging mutations) (Saccone et al., 2000; Blanchard and Lynch, 2000), protection from chemical mutagens (Allen and Raven,

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

1996; Wright et al., 2009) and replication errors (Itsara et al., 2014; Kennedy et al., 2013), and enhanced fixing of useful mutations (Adams and Palmer, 2003; Blanchard and Lynch, 2000). However, these observations raise the complementary question: why are any genes retained in organelles at all (Allen and Martin, 2016)? This question has been hotly debated over decades, with many proposed hypotheses. The preferential retention of genes encoding hydrophobic products has been suggested, due to the challenge of correctly targeting and importing such products to the correct organelle (Von Heijne, 1981; Popot and de Vitry, 1990; Björkholm et al., 2017). The retention of genes playing central roles in controlling redox processes has also been proposed to facilitate local, organellar control of activity (Allen, 2015). Other hypotheses, including roles for nucleic acid biochemistry (Johnston and Williams, 2016), gene expression levels (Nabholz et al., 2013), energetic costs of encoding (Kelly, 2021), toxicity (Martin and Schnarrenberger, 1997), and others have been proposed, but quantitative testing of these ideas remains limited (Johnston and Williams, 2016; Maciszewski and Karnkowska, 2019).

Applying tools from model selection to large-scale genomic data offers unprecedented and powerful opportunities to both generate and impartially test evolutionary and mechanistic hypotheses (Kirk et al., 2013) (aligning with an influential recent commentary on ideas in biology; Nurse, 2021). Here, following previous work on mitochondrial DNA (mtDNA) evolution (Johnston and Williams, 2016), we adopt this philosophy to explore the mechanisms shaping gene loss across organelles. First, mindful of the dangers of proposing parallels between different organelles (Smith and Keeling, 2015), we nonetheless hypothesized that the same genetic features would shape retention propensity of genes in mitochondrial and plastid DNA (ptDNA). Such features would predispose a gene to be more or less readily retained in oDNA overall, whereas the total extent of oDNA retention in a given species is shaped in parallel by functional and metabolic features (Maciszewski and Karnkowska, 2019; Hadariová et al., 2018) and evolutionary dynamics (characterized statistically in elegant recent work; Janouškovec et al., 2017). We further expect that these genetic features would reflect the above evolutionary tension, between maintaining genetic integrity and retaining the ability to obtain and control machinery, that applies to both organelles (Johnston, 2019; Adams and Palmer, 2003). With this general hypothesis in mind, we proceed by taking an impartial, data-driven approach using large-scale genomic data to investigate which physical, chemical, and genetic features of genes and their protein products predict oDNA gene retention. We will ask three linked questions: first, of the genes retained in oDNA by at least some eukaryotes, which features predict how commonly they are retained; second, which features distinguish organelle genes that are not retained in oDNA in any known eukaryotes; and third, which features predict the genes retained in more recent symbioses that resemble proto-organellar relationships.

RESULTS

Quantifying gene-specific oDNA loss patterns across eukaryotes

To quantitatively explore the features predicting oDNA gene retention, we first define a retention index for a given oDNA gene, measuring its propensity to be retained in oDNA. To this end, we acquired data on organelle gene content across eukaryotes, using 10,328 whole mtDNA and 5,176 whole ptDNA sequences from NCBI. We curated these data with two different approaches, resembling supervised and unsupervised philosophies, to form consistent records of gene presence/absence by species (see STAR Methods). The supervised approach (manual assignment of ambiguous gene records to a chosen gene label) and the unsupervised approach (all-against-all BLAST comparison of every gene record from the organelle genome database) agreed tightly (Figure S1).

Simply counting observations of each gene across species is prone to large sampling bias, as some taxa (notably bilaterians and angiosperms) are much more densely sampled than others. Instead, we reconstructed gene loss events using oDNA sequences of modern organisms and an estimated taxonomic relationship between them (see STAR Methods). We then define a retention index for each gene, quantifying its relative propensity to be retained in oDNA while accounting for the underlying phylogenetic connections between samples. Following the picture of hypercubic transition path sampling (Johnston and Williams, 2016; Greenbury et al., 2020), we use an evolutionary model representing all possible patterns of gene presence and absence as nodes on a directed hypercubic transition graph. Each node has a corresponding unique binary string label of length L, where a 0 or 1 in position i corresponds to absence or presence of the ith gene, respectively. Under this evolutionary model, the oDNA complement of species evolves by traversing edges from an ancestral state with all genes (the binary string of all 1s), progressively losing genes (and corresponding acquiring 0s in its state). This traversal occurs in parallel to phylogenetic branching, so that each phylogenetic lineage inherits its ancestral oDNA pattern and then traverses the hypercube independently thereafter. In this picture, we can define the sampled retention index of gene X as the mean number of other genes already lost when an inferred transition on the hypercube involves the loss of gene X. In other words, the retention index estimates the average number of genes already lost in a lineage when gene X is lost (our results were robust with respect to alternative definitions; see below). This retention index, along with the unique patterns of oDNA gene presence/absence and their taxonomic distribution, are illustrated in Figure 1 (phylogenetic embedding in Figure S2).

The retention patterns of genes in mtDNA and ptDNA across eukaryotes show pronounced structure, agreeing with existing results (and arguing against a null hypothesis of random gene loss). The several-fold expansion of mtDNA in this study compared with (Johnston and Williams, 2016) preserves the same structure, with, for example, several *rpl* genes and *sdh [2-4]* commonly lost and *nad[1-6]*, *cox[1-3]* and *cytb* commonly retained. The highest mtDNA protein-coding gene counts appear in jakobid protists, with over 60 protein-coding genes in some species. Metazoan mtDNA gene patterns are mostly stable, with a common 13-protein gene profile shared by a large majority of taxa (including humans). Viridiplantae have more diverse mtDNA profiles, typically containing more protein-coding genes; fungi are also highly diverse, with some large and some highly reduced gene profiles—typically retaining fewer protein-coding genes than plants. Parasitic species, notably alveolates, contain very few protein-coding mtDNA genes. The ptDNA patterns display pronounced clustering, following

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

CellPress
OPEN ACCESS

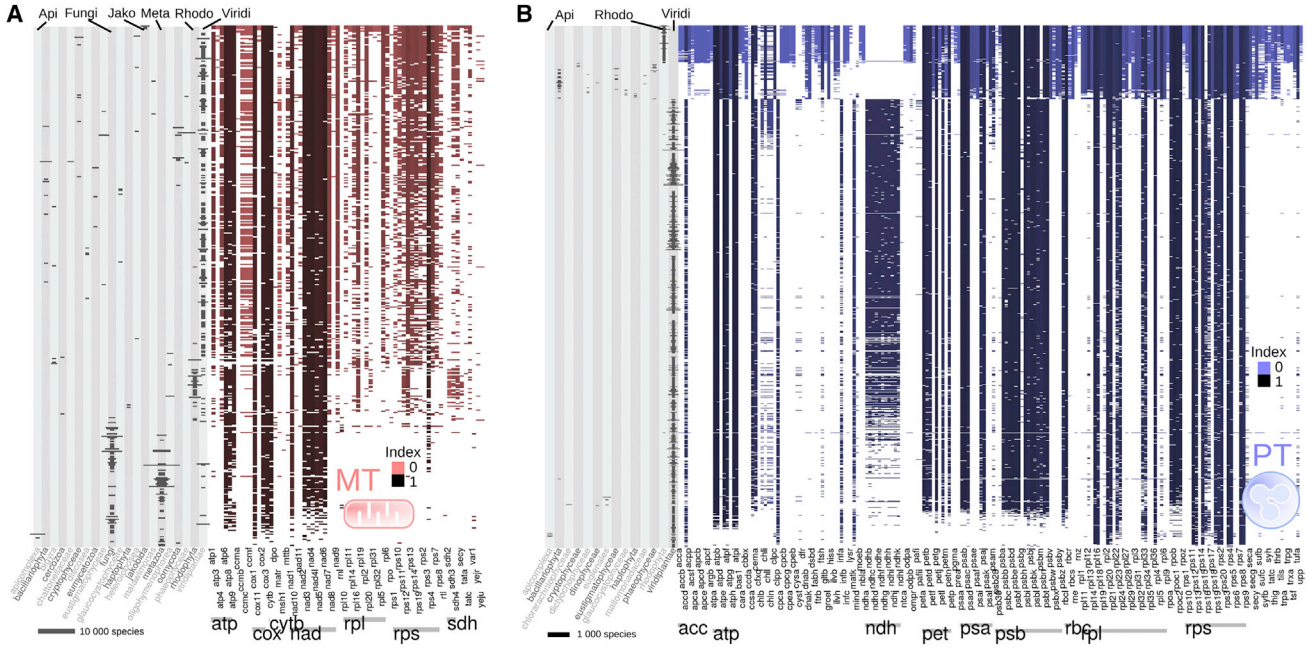


Figure 1. Structure of oDNA gene retention

Each row of colored/white pixels is a unique gene presence/absence pattern found in eukaryotic oDNA, where columns are individual oDNA genes in (A) mtDNA and (B) ptDNA. Darker colors correspond to higher values of our assigned retention index for a given gene. Each pattern may be present in many species: gray bars on the left of each row show the logarithm of the number of species with that pattern in a number of eukaryotic clades; scale bar below. The pronounced split in ptDNA patterns reflects the evolutionary pathways represented, for example, by Rhodophyta and Viridiplantae (Hohmann-Marriott and Blankenship, 2011). Sets of genes encoding subunits of notable organelle protein complexes are labeled with gray bars under the horizontal axis. Highlighted taxa with abbreviations are [Api]complexa, [Fungi], [Jako]bida, [Meta]zoa, [Rhodo]phyta, and [Viridi]plantae.

previous observations (Mohanta et al., 2020) and the known serial nature of their evolutionary heritage (Keeling, 2010), with one cluster corresponding broadly to Viridiplantae (typically retaining *ndh* genes) and the other corresponding broadly to brown and red algae, diatoms, and other clades (typically lacking *ndh* genes but retaining more *atp*, *rps*, *rpl*, *psa*, and *psb*). The largest ptDNA gene counts occur in the Rhodophyta, with the overall patterns of retention more distinctly clustered than for mtDNA. Viridiplantae, making up the majority of samples in the dataset, display a very large number of variations on a general structural theme, with many ptDNA gene sets shared by dozens of species. More reduced ptDNA gene sets are found in parasitic taxa, with the smallest found in vestigial plastids like apicoplasts in Apicomplexa. Several ribosomal subunits and *ndhb* are among the most retained in ptDNA, with a second tier involving many *ndh*, *psa*, *psb*, and *atp* genes retained in around half our species. Least retained ptDNA genes include other members of *psa*, *psb*, *rps*, and *rpl*.

Cross-organelle symmetry in the prediction of gene retention by hydrophobicity and GC content

To facilitate the principled investigation of different existing and new hypotheses on the features shaping oDNA gene retention, we proceed with a Bayesian model selection approach applied across oDNA. Here, the power of combinations of diverse molecular features to predict gene retention patterns are explored using the large-scale dataset alone, with no prior favoring of one hypothesis over another. To this end, we compiled a set of quantitative properties of genes and their protein products, linked to evolutionary hypotheses about the mechanisms shaping oDNA gene retention (Johnston and Williams, 2016). These included gene length and GC content, statistics of encoding and codon usage, protein hydrophobicity, molecular weight, energy requirements for production, average carboxyl and amino pK_a values for amino acid residues, and others (Figure S3). Our quantitative estimates for each feature were averages over a taxonomically diverse sampling of eukaryotic records (see STAR Methods). We used Bayesian model selection to ask which of these properties were most likely to be included in a linear model predicting the retention index of each gene. Following Johnston and Williams (2016), this approach identifies likely predictors with quantified uncertainty, while acting without prior favoring of any given hypotheses and automatically guarding against overfitting and the appearance of correlated predictors providing redundant information. In both mtDNA and ptDNA datasets, models where high hydrophobicity and high GC content predict high gene retention were strongly favored (Figure 2A). The model with the highest posterior probability that featured neither of these predictors instead featured amino pK_a and assembly energy, with a posterior probability of 3.14×10^{-4} —many orders of magnitude lower than the top models in Figure 2A, suggesting strong support for hydrophobicity and GC content as likely predictors. As mentioned above, the link with hydrophobicity has been discussed at length in the past. The link with GC content is, to our knowledge, less well-known—and requires a careful disambiguation. It is well-known that oDNA generally has lower GC content

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

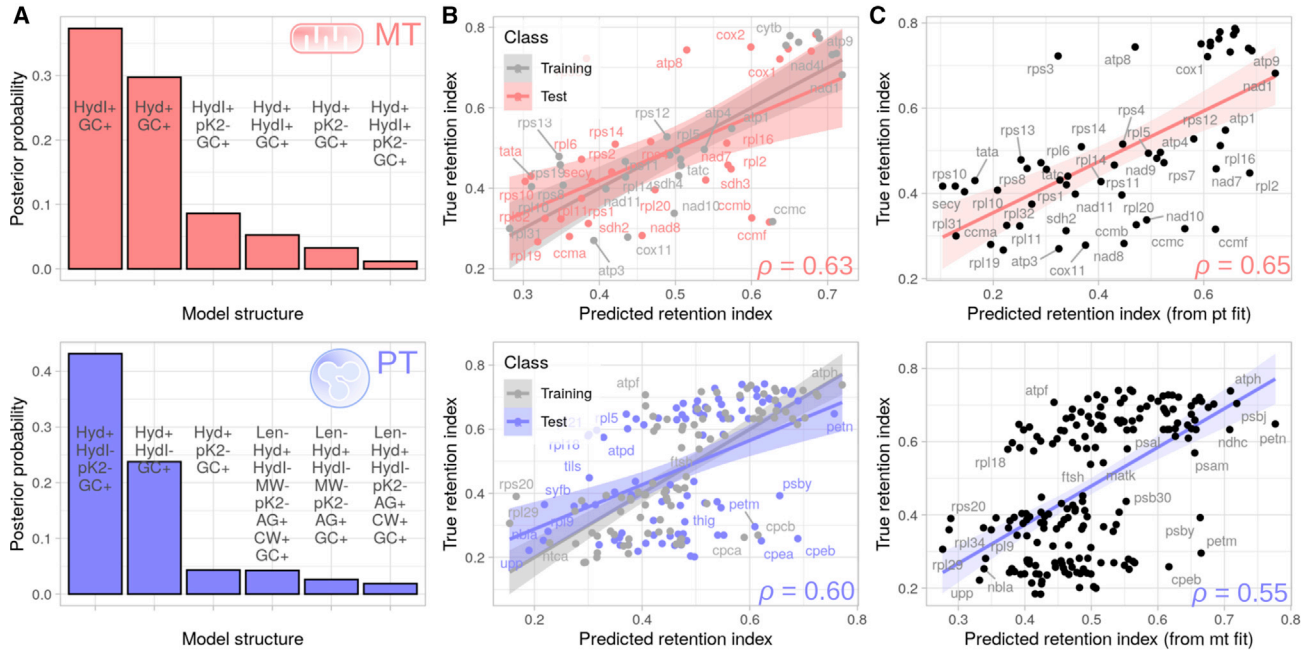


Figure 2. Predictors of oDNA gene retention

(A) Posterior probabilities over the set of features in linear models predicting retention index. Each model structure is given by a set of codes describing its component features. Hydrophobicity (Hyd) or hydrophobicity index (HydI) and GC content (GC) feature in all model structures with the highest posterior probabilities (for priors see STAR Methods). ±give posterior mean signs of associated coefficients in model for retention index. Abbreviations are as follows: Hyd, hydrophobicity; HydI, hydrophobicity index; GC, GC content; Len, length; pK1, carboxyl pK_a; pK2, amino pK_a; MW, molecular weight; and AG/CW, energies of gene expression (see STAR Methods).

(B) Prediction of retention index with linear models involving hydrophobicity and GC content. oDNA gene sets are split into training and test sets; trained models predict retention indices well in the independent test sets.

(C) Cross-organelle prediction. Linear models trained on mtDNA gene properties predict retention indices of ptDNA genes well, and vice versa. In (B) and (C), shaded regions give 95% confidence intervals.

than nuclear DNA because of the asymmetric mutational pressure arising from the hydrolytic deamination of cytosine to uracil, reducing GC content more in the generally high mutation compartments of oDNA than in the nucleus (Reyes et al., 1998). However, our results show that higher GC content is relatively favored between oDNA genes—and hence independently of the general oDNA/nDNA difference (Johnston and Williams, 2016).

We next tested the capacity of models involving this combination of features to predict the retention index of oDNA genes in an unseen dataset. We split mtDNA and ptDNA gene sets into 50:50 training and test sets, trained linear models involving hydrophobicity and GC content using the training data, and examined their performance in predicting retention index in the independent test set. Average Spearman correlations were $\rho = 0.64$ and $\rho = 0.62$ for training mt and pt sets, respectively, and $\rho = 0.63$ and $\rho = 0.60$ for test mt and pt sets, respectively (Figure 2B). Correlations were higher still ($\rho > 0.7$) when only subunits of core bioenergetic complexes were considered (Table S1). Following our hypothesis that the same features predict retention in the two organelle types, we also performed cross-organelle experiments. That is, we trained a hydrophobicity and GC model using mt genes and examined its ability to predict pt gene retention, and vice versa. Notably, both organelle gene sets predicted well the other's retention patterns ($\rho = 0.65$ for pt predicting mt; $\rho = 0.55$ for mt predicting pt; Figure 2C; Table S1). In other words, a simple model trained only using mitochondrial gene data can predict the retention profile of plastid genes, and vice versa.

To relax the assumptions involved in this approach, including linear modeling, we paralleled this analysis with a range of other regression approaches from data science, including penalized regression and random forests, and using different definitions of retention index (STAR Methods; Figure S4). We generally observed hydrophobicity and GC content being selected as features with good predictive ability and the capacity to predict one oDNA type's behavior from the other, regardless of statistical approach taken (Table S2); pK_a values were also selected as informative features in some model types (see below).

Importantly, there is clear evidence for a *combination* of features having predictive power over retention patterns. Previous work has often highlighted one feature over others; the model selection process here (as with previous work on mtDNA; Johnston and Williams, 2016) emphasizes that several features likely act in concert to shape oDNA patterns. The construction and validation of predictive models further allow these individual contributions to be quantified, addressing the problem of multiple causes (Hilborn and Stearns, 1982) (and to some extent explaining the ongoing debate between single-factor explanations in the field).

Hydrophobicity and protein biochemistry predict oDNA gene transfer to the nucleus in both organelles

We next asked which properties predict which organelle protein-coding genes are universally transferred to the nucleus across all eukaryotes. To this end, we compiled sets of annotated nDNA and oDNA genes encoding subunits of bioenergetic protein complexes in organelles using a custom pattern matching algorithm and 308 eukaryotic whole genome records from NCBI (see STAR Methods) (Figure 3A). As expected, GC content in organelle-encoded genes was systematically lower than nuclear-encoded genes. Here, this signal cannot be regarded as a causal mechanism because it is likely due at least in part to the aforementioned differences in asymmetric mutational pressure between nDNA and oDNA—organelles, which often have higher mutation rates than the nucleus, experience more asymmetric mutation driving GC content lower (Reyes et al., 1998; Johnston and Williams, 2016). More interestingly, the hydrophobicity of organelle-encoded genes was systematically higher across taxa (agreeing with hypotheses above, and recent observations in the mitoribosome; Bertgen et al., 2020). Other less well-studied biochemical features, including the carboxyl pK_a values of protein products, were also systematically different between encoding compartments (Figures 3A, 3B, and S5). We also verified that these differences existed within the sets of genes encoding subunits of different organellar complexes (Figures 3B and S6).

Following the above philosophy of data-driven identification of features shaping encoding compartment, we used Bayesian model selection with a generalized linear model (GLM) using gene properties to predict the encoding compartment (except GC and codon use statistics, due to the possibility of differences therein arising simply due to asymmetric mutation). We found, not unexpectedly given previous work, that hydrophobicity consistently appeared in all the model structures with highest posterior probability (Figure 3C). More surprisingly, high-probability predictive models also featured the relatively under-examined feature of carboxyl pK_a, in combination with hydrophobicity. Their appearance together in a Bayesian model selection framework suggests that they provide independent information on gene encoding, despite a correlation (albeit rather weak) between the features (Figure S3).

We next asked if these features could predict the encoding compartment of genes in an unseen dataset. To this end, we constructed GLMs predicting encoding compartment using hydrophobicity and carboxyl pK_a. Here, we consider each species in our dataset separately to avoid having to account for phylogenetic correlations between species—these models then give species-level predictions about encoding compartment. We trained this model using a subset of genes from each given species and asked how well it predicted encoding compartment in the remaining set of genes in that organisms. The model was able to predict the encoding compartment of an independent test set from each species with high performance (true positive [TP]/negative [TN] rates: mt TP 0.90 ± 0.17, TN 0.97 ± 0.10, pt TP 0.75 ± 0.20, TN 0.88 ± 0.18, mean and SD across the different species considered; Figure 3D). We employed a range of classification approaches to quantify these observations, again training on a subset of the observations and testing classification performance on an independent set (Figure S7). Hydrophobicity and pK_a values consistently appeared as strong separating terms, with other features including production energy and gene length playing a supporting role (Figure S7). Classification accuracy was typically >0.8 for all complexes using random forest approaches (Table S3).

For a subset of organelle-localized gene products, solved crystal structures of their protein complexes allow another property to be quantified: the binding energy statistics of the protein product in its protein complex structure. Here, we estimate the free energy associated with each subunit-subunit interaction in the solved complex using PDBePISA (Velankar et al., 2009) (results were robust when additional ligands were ignored or were considered part of the subunit-subunit interaction, see STAR Methods). We then compute the total free energy across all of a subunit's individual interfaces (the strength of its interactions within the complex), as a continuous quantitative proxy for its potential control over the assembly and stability of complex structure (Levy et al., 2008; Maier et al., 2013) and hence for a complex's role in redox regulation, as discussed in Allen and Martin (2016).

Previous work qualitatively suggested that genes encoding subunits with high total binding energy (strong binding interactions with neighboring subunits) and playing central roles in complex assembly pathways were most retained in mtDNA (Johnston and Williams, 2016; Maier et al., 2013; Allen and Martin, 2016). We used a generalized linear mixed model to quantify and extend this analysis to complexes in both organelle types. We found that total binding energy predicted whether a gene was organelle encoded in any eukaryotes, with the relationship holding across mitochondria and plastids, although with varying magnitudes in different complexes (Figures 3E and S8). We verified the absence of pronounced correlation structure between binding energy statistics and hydrophobicity (Figure S9), suggesting that the two features independently contribute to gene retention (Johnston and Williams, 2016). Hence, hydrophobicity, amino acid biochemistry, and energetic centrality (linked to co-localization for redox regulation; Allen and Martin, 2016) predict whether a gene is ever retained in oDNA; of those that are hydrophobicity and GC content predict the extent of this retention across eukaryotes.

Independent endosymbiotic genomes show compatible profiles of hydrophobicity and protein biochemistry

Evolutionary history cannot easily be rerun to independently examine the principles predicted by our analysis. However, the diversity of eukaryotic life provides some existing opportunities to test them. In several eukaryotic species, unicellular endosymbionts that are not directly related to mitochondria or plastids have co-evolved with their "host" species, in many cases involving gene loss and in some cases transfer of genes to the host. Class *Insecta* is known to have several examples of reduced bacterial endosymbionts (Husnik and Keeling, 2019); other notable examples include the chromatophore, an originally cyanobacterial endosymbiont of *Paulinella* freshwater amoebae (Gabr et al., 2020), the recently discovered *Candidatus Azoamicus ciliaticola*, a denitrifying gammaproteobacterial endosymbiont within a *Plagiopylea* ciliate host (Graf et al., 2021), and the *Nostoc azollae* symbiont of the *Azolla* water ferns (Ran et al., 2010).

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

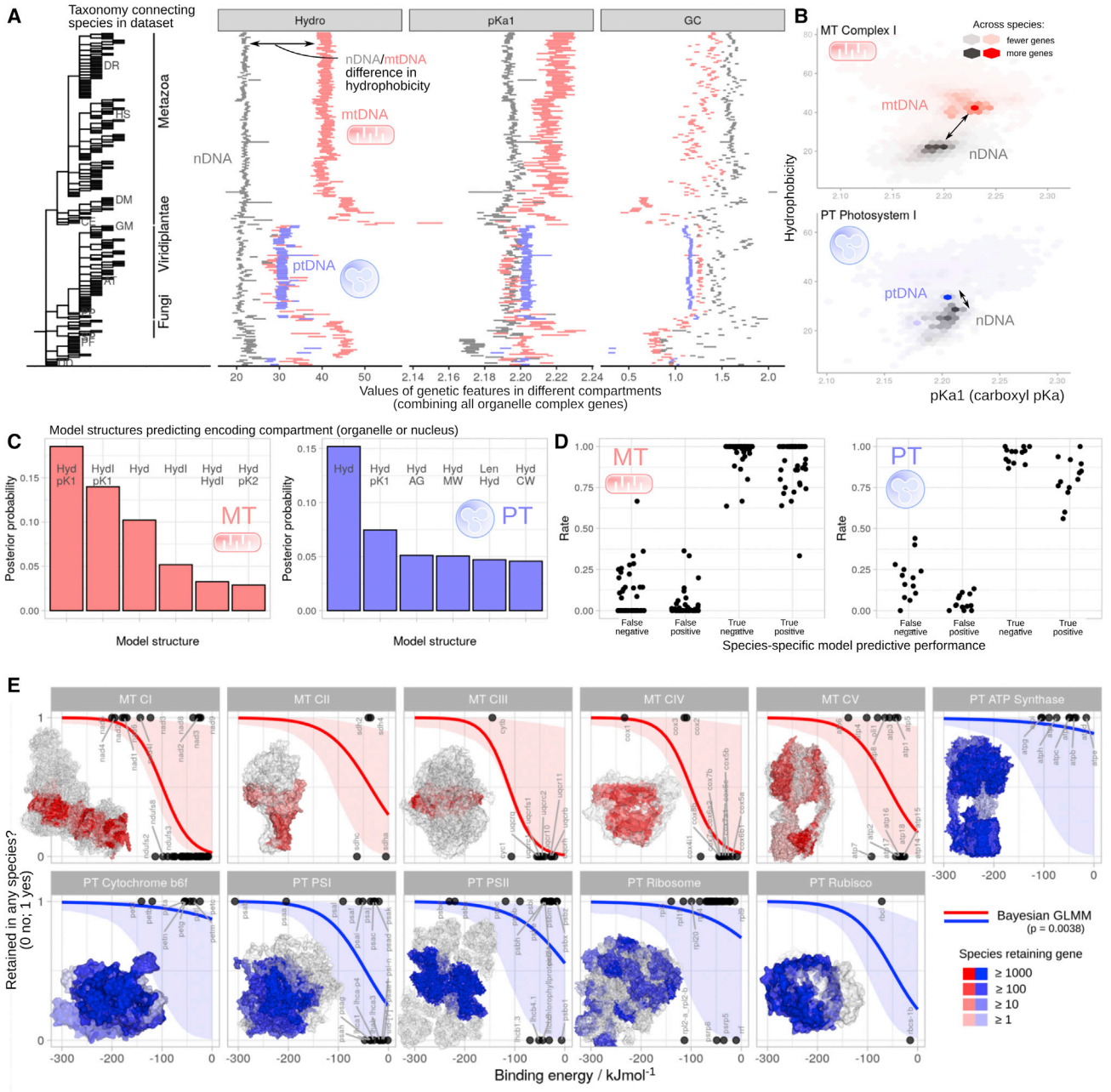


Figure 3. Features predicting encoding compartment

(A) Mean and SEM of selected gene properties for organelle genes encoded in nuclear DNA (gray), mtDNA (red), and ptDNA (blue), in different species (organized by the phylogeny on the left, expanded set in Figure S5).

(B) Hydrophobicity and carboxyl pK_a of organelle genes encoded in nuclear DNA (gray) and oDNA (red/blue), for two example protein complexes (expanded set in Figure S6). The darkness of a segment's color increases with the number of individual gene data points contained within it.

(C) Bayesian model selection with a generalized linear model (GLM) framework for features predicting the encoding compartment of a given gene. Posterior probabilities are averaged across independent classifications for individual organisms. Each model structure is given by a set of codes describing its component features; model labels as in Figure 2.

(D) Performance (T/F, true/false; P/N, positive/negative) of GLMs involving hydrophobicity and carboxyl pK_a on predicting encoding compartment of genes outside the training set. Each set of points corresponds to a model for one organism.

(E) Binding energy and encoding compartment. "Retention" here is not the continuous index used previous, but a binary variable describing if any organisms retain a gene in their oDNA (1) or if none have been observed to do so (0). Traces show mean and 95% credible intervals for Bayesian generalized linear mixed model (GLMM) (see STAR Methods for priors). The associated p value is a frequentist interpretation from bootstrapping, against the null hypothesis of no relationship. Crystal structures are colored according to the number of species in our dataset that retain the gene for each subunit.

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

CellPress
OPEN ACCESS

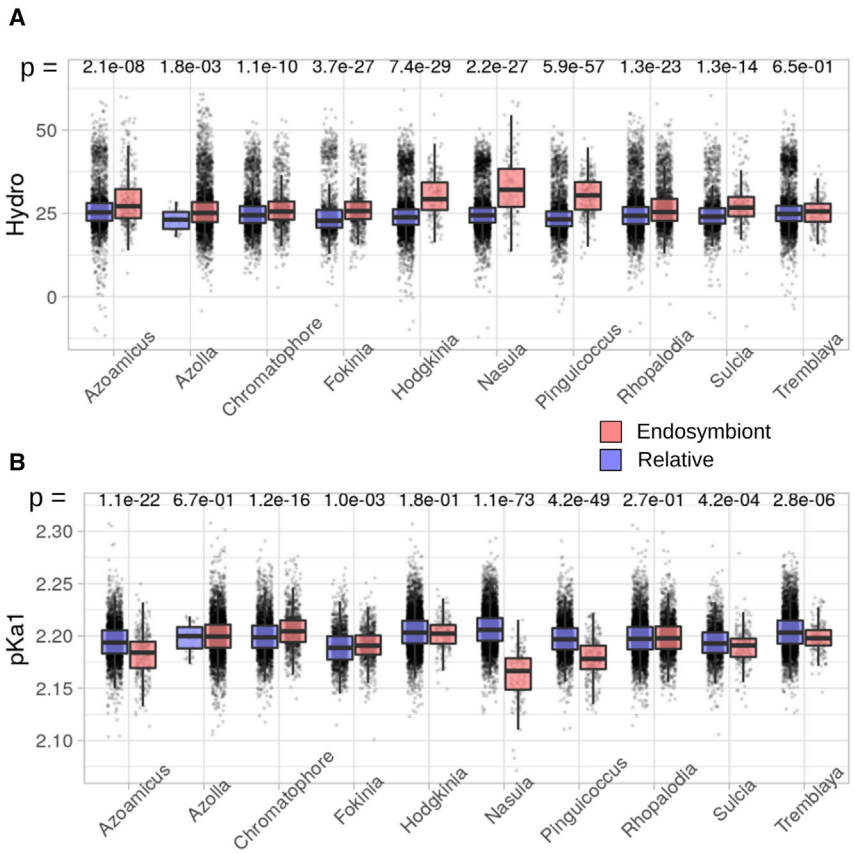


Figure 4. Gene feature profiles in other endosymbionts

(A) Hydrophobicity and (B) carboxyl pK_a across genes in endosymbionts (red) and a non-endosymbiotic close relative (blue). p values are from Wilcoxon rank-sum tests. Boxplots show upper and lower quartiles, medians, and whiskers to points not exceeding a factor of 1.5 from the upper and lower quartiles.

Not all of these endosymbiotic relationships have been shown to involve gene transfer to the host cell nucleus, although there is evidence for this in the *Paulinella* system (Nowack et al., 2011). The metabolic and energetic contexts, and ages, of the individual endosymbiotic relationships also differ, at least to some extent, in each case. A general principle can, however, be expected to hold across relationships. All cases involve reduction of the endosymbiont genome, as some machinery in the endosymbiont becomes redundant in the symbiotic relationship. In a subset of lost genes, this redundancy arises because host-encoded machinery can fulfill the required function (other genes will be lost without such host-encoded compensation, as their entire function becomes redundant). For this subset, the same broad principles regarding import of protein machinery may then be expected to hold as in organelles. Such genes are lost as host-encoded machinery removes the need for their local encoding. However, such host-encoded machinery must be physically acquired by the endosymbiont, raising similar issues of the mistargeting and import difficulty for hydrophobic gene products as in the organelle case. In tandem, any biochemical pressures influencing the ease of gene expression in the endosymbiont compartment may also be expected to shape retention patterns of this subset of genes. We therefore hypothesized that the principles we find to shape gene retention in mitochondria and plastids would also show a detectable signal in these independent endosymbiotic cases (while expecting a lower magnitude hydrophobicity signal, due to loss of some genes without the requirement for nuclear compensation).

To test this hypothesis, we computed genetic statistics for the genomes of endosymbionts and non-endosymbiotic close relatives (STAR Methods; Table S4). The hydrophobicity profile of the endosymbionts in 9 of 10 cases was significantly higher than their non-endosymbiotic relative (STAR Methods; Figure 4). Genes retained in the photosynthetic chromatophore also had lower carboxyl pK_a values than in a free-living relative; for other endosymbionts, this relationship was reversed, with endosymbiont genes having lower carboxyl pK_a values. This diversity is compatible with a possible mechanistic link between the pH of the compartment and the dynamics of gene expression therein (see discussion).

Our analysis approach involves several choices of parameter and protocol. To assess the robustness of our findings, we have varied these choices and checked the corresponding change in outputs, described in STAR Methods and the following figures. The key choices, with figures illustrating their effects, are in gene annotation (supervised or unsupervised; Figure S1), initial selection of features (where we followed existing hypotheses and particularly their summary in Johnston and Williams, 2016) and how to summarize their quantitative values (Figure S10), definition of retention index (Table S1;Figure S11), choice of priors in Bayesian model selection (Figure S12), and choice of regression and classification methods: we additionally tested least absolute shrinkage and selection operator (LASSO) and ridge regression and decision trees and random forests for regression and classification (Figures S7 and S11).

DISCUSSION

Present-day organelle genomes reflect a balance between pressure favoring gene retention and those favoring gene loss. Transfer to the nucleus is likely favored for protection of information (Adams and Palmer, 2003), with other features including the cellular ATP budget (Kelly, 2021) also playing a role. We have found, using a unique combination of large-scale genomic data, evolutionary modeling, and Bayesian model selection, that a constellation of previously and newly considered genetic features constitute opposing pressures to retain genes in organelles across eukaryotes, predicting gene retention to a notably symmetric extent in mitochondria, chloroplasts, and independent

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

endosymbionts. Protein product hydrophobicity, a common hypothesis, is confirmed as a highly likely predictor. However, it does not have sole predictive power over oDNA patterns. Our approach reveals quantitative roles for colocation for redox regulation (via the proxy of centrality of a subunit in its complex; Levy et al., 2008; Allen and Martin, 2016) and several newly considered features of nucleic acid and amino acid biochemistry (including GC content and carboxyl pK_a). It must be underlined that no single mechanism has sole predictive power over this behavior. As expected in complex biological systems, a combination of factors is likely at play, a situation that has perhaps contributed to the ongoing debate on this topic (Hilborn and Stearns, 1982)—and one which can be resolved through model selection approaches.

The cross-taxa structure of protein-coding gene profiles in mitochondria and plastids has been explored previously across many different phylogenetic scales; two broad recent examples are Johnston and Williams (2016) and Mohanta et al. (2020). There are some features of these patterns that should be mentioned in discussing our approach. The ptDNA profiles show a pronounced clustering into two classes, broadly reflecting the Viridiplantae and Rhodophyta "themes" (mirroring the evolutionary heritage of the organelles (Keeling, 2010)). This structure is reflected in the bimodal appearance of the plots in Figure 2 and is the reason for the slightly weaker predictive performance of plastid models than mitochondrial models throughout (Figures 2 and 3). Broadly, the taxa from which ptDNA profiles are drawn explain some amount of the variance in the data, but our "universal" models deliberately use information across eukaryotes without further phylogenetic subdivisions. Modeling the two ptDNA "themes" separately will certainly improve model performance, with the output being more taxa-specific models.

Binding energy predicts retention propensity most strongly at very low (under -200 kJ mol^{-1}) binding energies—very strongly bound subunits are almost invariably found in at least some eukaryotic oDNA. At higher energies (less strongly bound subunits), there is typically a mix of retention scores. This reflects a general observation—the interactions between features shaping oDNA gene retention can be complicated. At low energies, it may be that binding energy dominates retention behavior, whereas at higher energies, other features like hydrophobicity and pK_a play a competing role in determining retention. As we cannot analyze binding energy in the same model structure as our other features (the source datasets and response variables are different)—and as we use simple linear models for other features—the quantitative details of such interactions remain to be elucidated. A related observation is the cluster of mtDNA genes including *cox1*, *cytb*, and other very highly retained genes. These consistently appear above our model predictions—they are (slightly) more retained than our model predicts. This discrepancy may be accounted for by considering more flexible model structures or additional predictive genetic features.

Our connection with more recent endosymbionts also warrants some discussion. Although difficult to state precisely, the symbiotic relationships there span a range of ages—with estimates, for example, ranging from under 30 to 300 Ma for some insect symbioses (Moran and Wernegreen, 2000) and 90–140 Ma for *Paulinella* (Gabr et al., 2020). The different relationships have had different times to undergo evolution, and the symbionts and the free-living neighbors with which we paired them also have a range of divergence times (which we have not attempted to quantify, as data are limited for many pairs). There are several reasons that we may expect the strength of the effects we observe to differ between cases, but the direction of these effects does seem consistent enough to suggest a general theme. Further examples and data will help reveal how general this observation is and how strong an effect such pressures may provide over random drift (Boscaro et al., 2017).

How are evolutionary pressures from the features we identify manifest at the cellular level? Hydrophobic gene retention may be favored due to the physical difficulty of importing hydrophobic products or their propensity to be mistargeted to other compartments (Von Heijne, 1981; Björkholm et al., 2017) (although these mechanisms are not free from debate (Allen, 2015)). As a proxy for control over complex assembly (Levy et al., 2008; Maier et al., 2013) and thus regulation of the redox processes those complexes facilitate, the binding energy centrality of a subunit in its protein complex aligns with the CoRR (colocalization for redox regulation) hypothesis (Allen, 2015), where organelles retain genes that facilitate local redox control. GC content and carboxyl pK_a have less established mechanistic hypotheses and constitute an avenue for further research. The increased chemical stability of GC bonds (Samuels, 2005) has been suggested to support the integrity of oDNA in the damaging chemical environment of the organelle, meaning that high GC content (and hence more stable nucleic acid) may be favored in oDNA. pK_a, reflecting the ease of deprotonation of amino acid subgroups for different pH environments, influences the dynamics of peptide formation in translation (Watts and Forster, 2010), resulting in pronounced and diverse pH dependence of peptide formation for different amino acids (Johansson et al., 2011). Speculatively, we thus hypothesize that the synthesis of protein products enriched for higher-pK_a amino acids may involve lower kinetic hurdles in the more alkaline pH of mitochondria, plastids, and the chromatophore, favoring the retention of the corresponding genes. The pH within other endosymbionts, which perform less or no proton pumping, is expected to be lower, in which case the opposite pK_a trend observed in Figure 4 follows this pattern. This harnessing of large-scale sequence data with tools from model selection and machine learning has thus generated, and tested, new understanding of the fundamental evolutionary forces shaping bioenergetic organelles, providing quantitative support for several existing hypotheses and suggesting new contributory mechanisms to this important process.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS

- Source data
- Biochemical and biophysical properties of genes and products
- Binding energy calculations
- Pattern matching for nuclear-encoded organelle genes
- Gene labelling and evolutionary transitions
- Retention indices
- Prediction of retention index
- Classification of subcellular encoding
- Endosymbionts and relatives

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Code and dependencies

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2022.08.007.

ACKNOWLEDGMENTS

L.R. and J.M.C. are 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 program (grant agreement no. 805046 [EvoConBiO] to I.G.J.).

AUTHOR CONTRIBUTIONS

Conceptualization: I.G.J.; methodology: K.G., S.J.A., L.R., S.G., and I.G.J.; software: K.G., S.J.A., L.R., S.G., and I.G.J.; validation: K.G. and I.G.J.; formal analysis: K.G. and I.G.J.; investigation: K.G., S.J.A., L.R., J.M.C., E.C.R., and I.G.J.; data curation: K.G., S.J.A., L.R., J.M.C., and I.G.J.; writing – original draft: I.G.J.; writing – review and editing: K.G., S.J.A., L.R., S.G., J.M.C., E.C.R., and I.G.J.; visualization: S.J.A., S.G., and I.G.J.; supervision: I.G.J.; project administration: I.G.J.; funding acquisition: I.G.J.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 1, 2022
Revised: June 24, 2022
Accepted: August 22, 2022
Published: September 16, 2022

SUPPORTING CITATIONS

Enomoto et al., 2017; Floriano et al., 2018; Lhee et al., 2019; McCutcheon et al., 2009; McCutcheon and Moran, 2007; Nakayama and Inagaki, 2017; Serra et al., 2020; Bennett and Moran, 2013

REFERENCES

Adams, K.L., and Palmer, J.D. (2003). Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol. Phylogenet. Evol. *29*, 380–395.

Allen, J.F. (2015). Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. Proc. Natl. Acad. Sci. USA *112*, 10231–10238.

Allen, J.F., and Martin, W.F. (2016). Why have organelles retained genomes? Cell Syst. *2*, 70–72.

Allen, J.F., and Raven, J.A. (1996). Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J. Mol. Evol. *42*, 482–492.

Auguie, B. (2017). gridExtra: miscellaneous functions for "grid" graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Barton, M.D., Delneri, D., Oliver, S.G., Rattray, M., and Bergman, C.M. (2010). Evolutionary systems biology of amino acid biosynthetic cost in yeast. PLoS One *5*, e11935.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Software *67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bennett, G.M., and Moran, N.A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome Biol. Evol. *5*, 1675–1688.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242.

Bertgen, L., Mühlhaus, T., and Herrmann, J.M. (2020). Clingy genes: why were genes for ribosomal proteins retained in many mitochondrial genomes? Biochim. Biophys. Acta Bioenerg. *1861*, 148275.

Björkholm, P., Ernst, A.M., Hagström, E., and Andersson, S.G. (2017). Why mitochondria need a genome revisited. FEBS Lett. *591*, 65–75.

Blanchard, J.L., and Lynch, M. (2000). Organellar genes: why do they end up in the nucleus? Trends Genet. *16*, 315–320.

Booth, A., and Doolittle, W.F. (2015). Eukaryogenesis, how special really? Proc. Natl. Acad. Sci. USA *112*, 10278–10285.

Boscaro, V., Kolisko, M., Felletti, M., Vannini, C., Lynn, D.H., and Keeling, P.J. (2017). Parallel genome reduction in symbionts descended from closely related free-living bacteria. Nat. Ecol. Evol. *1*, 1160–1167.

Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A., and Andersson, S.G. (2004). Computational inference of scenarios for α-proteobacterial genome evolution. Proc. Natl. Acad. Sci. USA *101*, 9722–9727.

Bullerwell, C.E. (2011). Organelle Genetics: Evolution of Organelle Genomes and Gene Expression (Springer Science & Business Media).

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC bioinformatics *10* (1), 1–9.

Campitelli, E. (2021). ggnewscale: multiple fill and colour scales in 'ggplot2'. R package version 0.4.5. https://CRAN.R-project.org/package=ggnewscale.

Carr, D. (2021). hexbin: hexagonal binning routines. R package, version 1.28.2. https://CRAN.R-project.org/package=hexbin.

Cheng, J. (2021). ggpval: annotate statistical tests for 'ggplot2'. R package version 0.2.4. https://github.com/s6juncheng/ggpval.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika *78*, 685–709. https://doi.org/10.1007/s11336-013-9328-2.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J.L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Craig, C.L., and Weber, R.S. (1998). Selection costs of amino acid substitutions in cole1 and colia gene clusters harbored by *Escherichia coli*. Mol. Biol. Evol. *15*, 774–776.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems *1695*, 1–9. https://igraph.org.

DeLano, W.L. (2002). Pymol: an open-source molecular graphics tool. CCP4 Newsletter Pro. Crystallogr. *40*, 82–92.

Edwards, D.M., Røyrvik, E.C., Chustecki, J.M., Giannakis, K., Glastad, R.C., Radzvilavicius, A.L., and Johnston, I.G. (2021). Avoiding organelle mutational meltdown across eukaryotes with or without a germline bottleneck. PLOS Biol. *19*, e3001153.

Enomoto, S., Chari, A., Clayton, A.L., and Dale, C. (2017). Quorum sensing attenuates virulence in *Sodalis praecaptivus*. Cell Host Microbe *21*, 629–636.e5.

Federhen, S. (2012). The NCBI taxonomy database. Nucleic Acids Res. *40*, D136–D143.

Floriano, A.M., Castelli, M., Krenek, S., Berendonk, T.U., Bazzocchi, C., Petroni, G., and Sassera, D. (2018). The genome sequence of "*Candidatus* Fokinia solitaria": insights on reductive evolution in *Rickettsiales*. Genome Biol. Evol. *10*, 1120–1126.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22. https://www.jstatsoft.org/v33/i01/.

Gabr, A., Grossman, A.R., and Bhattacharya, D. (2020). Paulinella, a model for understanding plastid primary endosymbiosis. J. Phycol. *56*, 837–843.

Gelman, A., and Su, Y.S. (2020). arm: data analysis using regression and multilevel/hierarchical models. R package version 1.11-2. https://CRAN.R-project.org/package=arm.

Graf, J.S., Schorn, S., Kitzinger, K., Ahmerkamp, S., Woehle, C., Huettel, B., Schubert, C.J., Kuypers, M.M.M., and Milucka, J. (2021). Anaerobic endosymbiont generates energy for ciliate host by denitrification. Nature *591*, 445–450.

Greenbury, S.F., Barahona, M., and Johnston, I.G. (2020). Hypertraps: inferring probabilistic patterns of trait acquisition in evolutionary and disease progression pathways. Cell Syst. *10*, 39–51.e10.

Hadariová, L., Vesteg, M., Hampl, V., and Krajčovič, J. (2018). Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. Curr. Genet. *64*, 365–387.

Heinze, G., Ploner, M., and Jiricka, L. (2020). logistf: Firth's bias-reduced logistic regression. R package version 1.24. https://CRAN.R-project.org/package=logistf.

Hilborn, R., and Stearns, S.C. (1982). On inference in ecology and evolutionary biology: the problem of multiple causes. Acta Biotheor. *31*, 145–164.

Hohmann-Marriott, M.F., and Blankenship, R.E. (2011). Evolution of photosynthesis. Annu. Rev. Plant Biol. *62*, 515–548.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular biology and evolution *33* (6), 1635–1638.

Husnik, F., and Keeling, P.J. (2019). The fate of obligate endosymbionts: reduction, integration, or extinction. Curr. Opin. Genet. Dev. *58–59*, 1–8.

Itsara, L.S., Kennedy, S.R., Fox, E.J., Yu, S., Hewitt, J.J., Sanchez-Contreras, M., Cardozo-Pelaez, F., and Pallanck, L.J. (2014). Oxidative stress is not a major contributor to somatic mitochondrial dna mutations. PLoS Genet. *10*, e1003974.

Janouškovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Rohwer, F.L., Mylnikov, A.P., and Keeling, P.J. (2017). A new lineage of eukaryotes illuminates early mitochondrial genome reduction. Curr. Biol. *27*, 3717–3724.e5.

Johansson, M., Ieong, K.W., Trobro, S., Strazewski, P., Åqvist, J., Pavlov, M.Y., and Ehrenberg, M. (2011). ph-sensitivity of the ribosomal peptidyl transfer reaction dependent on the identity of the a-site aminoacyl-tRNA. Proc. Natl. Acad. Sci. USA *108*, 79–84.

Johnson, V.E., and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. J. R. Stat. Soc. B *72*, 143–170.

Johnston, I.G. (2019). Tension and resolution: dynamic, evolving populations of organelle genomes within plant cells. Mol. Plant *12*, 764–783.

Johnston, I.G., and Burgstaller, J.P. (2019). Evolving mtdna populations within cells. Biochem. Soc. Trans. *47*, 1367–1382.

Johnston, I.G., and Williams, B.P. (2016). Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. Cell Syst. *2*, 101–111.

Kassambara, A. (2020). ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr.

Keeling, P.J. (2010). The endosymbiotic origin, diversification and fate of plastids. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 729–748.

Kelly, S. (2021). The economics of organellar gene loss and endosymbiotic gene transfer. Genome Biol. *22*, 345.

Kennedy, S.R., Salk, J.J., Schmitt, M.W., and Loeb, L.A. (2013). Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. *9*, e1003794.

Kirk, P., Thorne, T., and Stumpf, M.P. (2013). Model selection in systems and synthetic biology. Curr. Opin. Biotechnol. *24*, 767–774.

Lane, N., and Martin, W. (2010). The energetics of genome complexity. Nature *467*, 929–934.

Levy, E.D., Boeri Erba, E.B., Robinson, C.V., and Teichmann, S.A. (2008). Assembly reflects evolution of protein complexes. Nature *453*, 1262–1265.

Lhee, D., Ha, J.S., Kim, S., Park, M.G., Bhattacharya, D., and Yoon, H.S. (2019). Evolutionary dynamics of the chromatophore genome in three photosynthetic Paulinella species. Sci. Rep. *9*, 2560.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. R J. *2*, 18–22. https://CRAN.R-project.org/doc/Rnews/.

Lide, D. (1991). Handbook of Chemistry and Physics (CRC Press).

Maciszewski, K., and Karnkowska, A. (2019). Should I stay or should I go? retention and loss of components in vestigial endosymbiotic organelles. Curr. Opin. Genet. Dev. *58–59*, 33–39.

Maier, U.G., Zauner, S., Woehle, C., Bolte, K., Hempel, F., Allen, J.F., and Martin, W.F. (2013). Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. Genome Biol. Evol. *5*, 2318–2329.

Martin, W., and Schnarrenberger, C. (1997). The evolution of the calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Curr. Genet. *32*, 1–18.

Martin, W.F., Garg, S., and Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370* 20140330.

McCutcheon, J.P., McDonald, B.R., and Moran, N.A. (2009). Origin of an alternative genetic code in the extremely small and gc–rich genome of a bacterial symbiont. PLoS Genet. *5*, e1000565.

McCutcheon, J.P., and Moran, N.A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl. Acad. Sci. USA *104*, 19392–19397.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2021). e1071. Misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package, version 1.7-8. https://CRAN.R-project.org/package=e1071.

Mohanta, T.K., Mishra, A.K., Khan, A., Hashem, A., Abd_Allah, E.F., and Al-Harrasi, A. (2020). Gene loss and evolution of the plastome. Genes *11*, 1133.

Monera, O.D., Sereda, T.J., Zhou, N.E., Kay, C.M., and Hodges, R.S. (1995). Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. J. Pept. Sci. *1*, 319–329.

Moran, N.A., and Wernegreen, J.J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. *15*, 321–326.

Nabholz, B., Ellegren, H., and Wolf, J.B. (2013). High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. Mol. Biol. Evol. *30*, 272–284.

Nakayama, T., and Inagaki, Y. (2017). Genomic divergence within non-photosynthetic cyanobacterial endosymbionts in rhopalodiacean diatoms. Sci. Rep. *7*, 13075.

Nowack, E.C., Vogel, H., Groth, M., Grossman, A.R., Melkonian, M., and Glöckner, G. (2011). Endosymbiotic gene transfer and transcriptional regulation of transferred genes in Paulinella chromatophora. Mol. Biol. Evol. *28*, 407–422.

Nurse, P. (2021). Biology must generate ideas as well as data. Nature *597*, 305.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., and Pearse, W. (2018). caper: comparative analyses of phylogenetics and

evolution in r. R package version 1.0.1. https://CRAN.R-project.org/package=caper.

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528.

Pennell, M.W., Eastman, J.M., Slater, G.J., Brown, J.W., Uyeda, J.C., Fitzjohn, R.G., Alfaro, M.E., and Harmon, L.J. (2014). geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics *30*, 2216–2218.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team. (2021). nlme: linear and nonlinear mixed effects models. R package version 3.1-152. https://CRAN.R-project.org/package=nlme.

Popot, J.L., and de Vitry, C. (1990). On the microassembly of integral membrane proteins. Annu. Rev. Biophys. Biophys. Chem. *19*, 369–403.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K.Y. (2021). BMA: bayesian model averaging. R package version 3.18.15. https://CRAN.R-project.org/package=BMA.

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J.A., Ininbergs, K., Zheng, W.W., Lapidus, A., Lowry, S., Haselkorn, R., and Bergman, B. (2010). Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One *5*, e11486.

Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223.

Reyes, A., Gissi, C., Pesole, G., and Saccone, C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. *15*, 957–966.

Ripley, B. (2021). tree: classification and regression trees. R package version 1.0-41. https://CRAN.R-project.org/package=tree.

Rossell, D., Cook, J.D., Telesca, D., Roebuck, P., and Abril, O. (2021). mombf: Bayesian model selection and averaging for non-local and local priors. R package version 3.0.4. https://CRAN.R-project.org/package=mombf.

Saccone, C., Gissi, C., Lanave, C., Larizza, A., Pesole, G., and Reyes, A. (2000). Evolution of the mitochondrial genetic system: an overview. Gene *261*, 153–159.

Samuels, D.C. (2005). Life span is related to the free energy of mitochondrial dna. Mech. Ageing Dev. *126*, 1123–1129.

Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592–593. https://doi.org/10.1093/bioinformatics/btq706.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Crowley, J. (2021). GGally: extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally.

Serra, V., Gammuto, L., Nitla, V., Castelli, M., Lanzoni, O., Sassera, D., Bandi, C., Sandeep, B.V., Verni, F., Modeo, L., and Petroni, G. (2020). Morphology, ultrastructure, genomics, and phylogeny of *Euplotes vanleeuwenhoeki* sp. nov. and its ultra-reduced endosymbiont "*Candidatus* Pinguicoccus supinus" sp. nov. Sci. Rep. *10*, 20311.

Sloan, D.B., Warren, J.M., Williams, A.M., Wu, Z., Abdel-Ghany, S.E., Chicco, A.J., and Havird, J.C. (2018). Cytonuclear integration and co-evolution. Nat. Rev. Genet. *19*, 635–648.

Slowikowski, K. (2021). ggrepel: automatically position non-overlapping text labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel.

Smith, D.R., and Keeling, P.J. (2015). Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. USA *112*, 10177–10184.

van der Loo, M.J., (2014). The stringdist package for approximate string matching. R J. *6*, 111–122. https://CRAN.R-project.org/package=stringdist.

Velankar, S., Best, C., Beuth, B., Boutselakis, C., Cobley, N., Da Silva, A.S., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M., et al. (2009). PDBe: protein data bank in Europe. Nucleic Acids Res. *39*, D402–D410.

Von Heijne, G. (1981). On the hydrophobic nature of signal sequences. Eur. J. Biochem. *116*, 419–422.

Watts, R.E., and Forster, A.C. (2010). Chemical models of peptide formation in translation. Biochemistry *49*, 2177–2185.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York). https://ggplot2.tidyverse.org.

Wickham, H. (2019). stringr: simple, consistent wrappers for common string operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr.

Wilke, C.O. (2020). cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot.

Wright, A.F., Murphy, M.P., and Turnbull, D.M. (2009). Do organellar genomes function as long-term redox damage sensors? Trends Genet. *25*, 253–261.

Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L., Tang, W., Feng, T., Chen, M., Zhan, L., et al. (2021). ggtreeextra: compact visualization of richly annotated phylogenetic data. Mol. Biol. Evol. *38*, 4039–4042.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. *8*, 28–36. https://doi.org/10.1111/2041-210X.12628. http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12628/abstract.

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
NCBI RefSeq Organelle Sequences	O'Leary et al., 2016	https://www.ncbi.nlm.nih.gov/genome/organelle/ (all organelle genomes available)
NCBI Whole Genome Assemblies		https://www.ncbi.nlm.nih.gov/genome/ (all eukaryotic genomes available)
NCBI Taxonomy Common Tree	Federhen, 2012	https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi
PDB Structures	Berman et al., 2000	https://www.rcsb.org/ PDB: 1oco, 1q90, 2h88, 2wsc, 5iu0, 5mdx, 5mlc, 5o31, 5xte, 6cp3, 6fkf.
Software and algorithms		
Full pipeline	This paper	https://github.com/StochasticBiology/odna-loss https://doi.org/10.5281/zenodo.6873510
PDBePISA	Velankar et al., (2009)	https://www.ebi.ac.uk/pdbe/pisa/
BLAST	Camacho et al., 2009	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
R	R Core Team, 2021	https://www.r-project.org/
ape (R library)	Paradis and Schliep, 2019	(install within R)
arm (R library)	Gelman and Su, 2020	(install within R)
blme (R library)	Chung et al., 2013	(install within R)
BMA (R library)	Raftery et al., 2021	(install within R)
caper (R library)	Orme et al., 2018	(install within R)
cowplot (R library)	Wilke, 2020	(install within R)
e1071 (R library)	Meyer et al., 2021	(install within R)
geiger (R library)	Pennell et al., 2014	(install within R)
GGally (R library)	Schloerke et al., 2021	(install within R)
ggnewscale (R library)	Campitelli, 2021	(install within R)
ggplot2 (R library)	Wickham, 2016	(install within R)
ggpubr (R library)	Kassambara, 2020	(install within R)
ggpval (R library)	Cheng, 2021	(install within R)
ggrepel (R library)	Slowikowski, 2021	(install within R)
ggtree (R library)	Yu et al., 2017	(install within R)
ggtreeextra (R library)	Xu et al., 2021	(install within R)
glmnet (R library)	Friedman et al., 2010	(install within R)
gridExtra (R library)	Auguie, 2017	(install within R)
hexbin (R library)	Carr et al., 2021	(install within R)
igraph (R library)	Csardi and Nepusz, 2006	(install within R)
lme4 (R library)	Bates et al., 2015	(install within R)
logistf (R library)	Heinze et al., 2020	(install within R)
mombf (R library)	Rossell et al., 2021	(install within R)
nlme (R library)	Pinheiro et al., 2021	(install within R)
phangorn (R library)	Schliep, 2011	(install within R)
phytools (R library)	Revell, 2012	(install within R)
randomForest (R library)	Liaw and Wiener, 2002	(install within R)
stringdist (R library)	van der Loo, 2014	(install within R)
stringr (R library)	Wickham, 2019	(install within R)
tree (R library)	Ripley, 2021	(install within R)

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biopython (Python library)	Cock et al., 2009	https://biopython.org/
ETE3 (Python library)	Huerta-Cepas et al., 2016	(install within Python)
PyMOL	DeLano, 2002	https://pymol.org/2/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Iain Johnston (iain.johnston@uib.no).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.
- Code is written in R, Python, and C, with a wrapper script for bash, and is freely available at github.com/StochasticBiology/odna-loss. This study did not generate new unique reagents.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Source data

We used the mitochondrion and plastid sequences available from NCBI RefSeq (O'Leary et al., 2016), and annotated eukaryotic whole genome data also from NCBI. The accessions and references for the endosymbiont/relative pairs are given in Table S4. For biochemical and biophysical gene properties, we used the values from (Johnston and Williams, 2016), described below, using BioPython (Cock et al., 2009) to assign these to given gene sequences. We averaged gene statistics over representative species from a collection of diverse taxa, both using model species (*Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Reclinomonas americana*, *Chondrus crispus*, *Plasmodium falciparum*) and randomly selected members of different taxa (Figure S10). Codes used in the figures are [Hyd]rophobicity, [HydI] hydrophobicity index, [GC] content, [Len]gth, [pK1] carboxyl pK_a, [pK2] amino pK_a, [MW] molecular weight, [AG/CW] energies of gene expression. We used crystal structures and associated HTML descriptions from the PDB (Berman et al., 2000) references PDB: 1oco, 1q90, 2h88, 2wsc, 5iu0, 5mdx, 5mlc, 5o31, 5xte, 6cp3, 6fkf. We used PDBePISA (Velankar et al., 2009) to estimate subunit binding energies with two different protocols, both removing ligands and incorporating them into the overall binding energy value for a subunit. We used estimated taxonomies from NCBI's Common Taxonomy Tree tool (Federhen, 2012).

Biochemical and biophysical properties of genes and products

Our assignment of biochemical and biophysical properties of genes and their products follows that in Johnston and Williams (2016). The length* (in number of amino acids of gene product) and GC content (trivially counted) of genes are taken straightforwardly from a sequence. Chemical properties of amino acids were taken from the compilation at http://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html. The hydrophobicity and hydrophobicity index of a gene product was computed using this compilation (original data from Ref. Monera et al., 1995). Amine group pK_a, carboxyl group pK_a, and molecular weight* values were calculated using this compilation (original data from Lide, 1991).

Glucose energy costs* were computed using the $A_{glucose}$ metric, based on the absolute nutrient cost required for amino acid biosynthesis, from Ref. Barton et al., 2010. Craig-Weber energy costs*, estimating the number of high-energy phosphate bonds and reducing hydrogen atoms required from the cellular energy pool to produce an amino acid, were taken from Ref. Craig and Weber, 1998. These biochemical properties are summarised in Table S5.

Asterisks denote properties that are not averaged over gene length; it was deemed more appropriate to average other properties over genome length to gain a representative measure. To check for artefacts from this interpretation, we performed a (much more computationally demanding) model selection process including both the normalised and un-normalised values for each property; although coverage of individual models was unavoidably low in this procedure, the same consistent observation of GC content and hydrophobicity as important features was observed throughout.

To compute a single value for each statistic of interest, a protocol is required to summarise the many different values seen for a given gene across the species in our dataset. For robustness, we considered several different averaging protocols. First, we averaged gene statistics over a set of model species taken from diverse eukaryotic groups (*Homo sapiens*, *Arabidopsis thaliana*,

Please cite this article in press as: Giannakis et al., Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention, Cell Systems (2022), https://doi.org/10.1016/j.cels.2022.08.007

Saccharomyces cerevisiae, *Reclinomonas americana*, *Chondrus crispus*, *Plasmodium falciparum*). Second, we randomly selected a member of each clade branching from the eukaryotic group (see clade names above) and averaged over the set containing these random samples. Most statistics were very strongly correlated for these different choices (Figure S10A). The exception was GC content, which is well known to evolve differently in different clades. To assess the effect of this difference, we ran the model selection process in the text with randomly-sampled averaging protocols. We found that despite differences in GC statistics, the selected models, and the presence of GC within them, remained robust to averaging choice (Figure S10B).

Binding energy calculations

We used PDBePISA (Velankar et al., 2009) to calculate interaction energies between different protein subunits and ligands in crystal structures. We summed the interaction energies over all interfaces between a given subunit and its partners to compute a total energetic centrality statistic for each subunit. Several choices of representation are possible for these calculations. Ligands can be ignored, so that only interaction energies of interfaces directly linking protein subunits are considered. Alternatively, bonds to ligands can be included as contributing to a given subunit's total binding energy. We primarily considered the mean energy per interface, including ligands, for each subunit, but also verified that our detected relationship existed for different choices including total energy over interfaces.

Pattern matching for nuclear-encoded organelle genes

Due to the previously mentioned inconsistency of annotation across species, we used a combination of positive and negative pattern matching with regular expressions to identify annotations for genes encoding subunits of different organelle complexes. After substantial preliminary experimentation to find good capture performance on a subset of the data, the positive matches required were:

CI	/NADH dehydrogenase|[Uu]biquinone oxidoreductase/
CII	/[Ss]uccinate dehydrogenase|[cC]o[qQ] reductase/
CIII	/[Cc]ytochrome [Bb]|[Cc]ytochrome [Cc] reductase/
CIV	/[Cc]ytochrome [cC] oxidase/
CV	/[Aa][Tt][Pp] synthase|ATPase sub/
MitoRibo	/[Rr]ibosomal.*[Mm]itochondri/
PSI	/[Pp]hotosystem I /
PSII	/[Pp]hotosystem II /
Cytb6f	/[Cc]ytochrome [Bb]6|[Cc]ytochrome f|[Pp]lastocyanin reductase/
Rubisco	/bi.phosphate [Cc]arboxylase/
PlastRibo	/[Rr]ibosomal.*[cC]hloroplast/

With the following patterns (split for formatting) required to be absent:
/assembly|alternative|containing|dependent|chaperone|kinase|NADH-cytochrome|coupling|maturase|
vacuolar|biogenesis|repair|LOW QUALITY PROTEIN|synthetase|activator|reticulum|activase|
synthesis|lyase|like| non|transporting|lipid|autoinhibited|membrane|type|required|
QUALITY|precursor|inhibitor|proteasomal|proteasome|E1|various|regulatory|Clp|
calcium|vesicle|b-245|b5|WRNIP|AAA|Cation|family|remodelling/

The outputs of this approach were manually verified to include genes encoding subunits physically present in their corresponding complex, while excluding assembly factors, regulatory factors, synthesis factors, unrelated enzymes, and other false positives.

Gene labelling and evolutionary transitions

Gene annotations are inconsistent across such a diverse dataset. For organelle genomes, we used two approaches. In a supervised approach, where the full set of unique labels found was manually curated and assigned a 'correct' label based on biological knowledge. In an unsupervised approach, we used BLASTn to perform an all-against-all comparison of all genes in our dataset. We scored each comparison as the proportional length of the region of identity compared to the reference sequence, multiplied by the proportion of identities across that region. Scores over 0.75 were interpreted as 'hits' (e.g. 75% identity over the full sequence, or full identity over 75% of the sequence). If more than 25% of appearance of gene label *X* in the BLAST output involved a 'hit' with gene labels *Y*, we interpreted *X* and *Y* as referring to the same gene. This process built a set of pairwise identities, which we then resolved iteratively into groups of gene labels assumed to refer to the same gene. We then assigned the most prevalent gene label to all members of that group. In each case, we retained only genes that were present in more than ten species in our dataset. For annotated whole genome data, we used pattern matching for gene annotations based on regular expression identifiers to identify nuclear-encoded subunits of organellar protein complexes (expressions below).

Using these curated gene sets, we assigned 'barcodes' of gene presence/absence (binary strings of length *L*, with 0 denoting gene absence and 1 denoting gene presence) to each species in our dataset. Each of these species is a tip on an estimated taxonomic tree

describing their putative evolutionary relationship. Assuming that gene loss is rare and gene gain is very rare, we iteratively reconstructed parent barcodes on this tree by assigning a 0 for gene *X* if all descendants lack *X*, and 1 otherwise. We then identified parent-child pairs where the child barcode had fewer genes than the parent (the opposite is impossible by construction). For each such instance, we record the transition from parent barcode to child barcode as a loss event.

Retention indices

Our simple retention index is defined as follows. Identify the set of transitions that involve the loss of gene *X*. For each transition in this set, count the genes retained by the parent and the genes retained by the child, and take their mean. The retention index is the mean of this quantity over the set of transitions where *X* is lost. The rationale is to characterise the number of genes that have already been lost when *X* is lost. If a transition event involves only the loss of *X*, the parent-child average will report this number minus $1/2$. If a transition involves the loss of several other genes in parallel with *X*, the average of the before and after counts is used.

In addition to our simple retention index, which relies on an estimated phylogeny linking observations in our dataset, we considered another assumption-free index. Here, we construct the set of unique oDNA presence/absence patterns in our dataset (as in Figure 1), and simply count the occurrences c_i of each gene i in this dataset. The index is given by $\log c_i / \max_j \log c_j$. This index relies on no evolutionary assumptions, and thus cannot account for the evolutionary relationship between sampled species. Considering only the set of unique barcodes goes some way towards accounting for the sampling bias in the dataset (for example, almost all metazoans have the same presence/absence profile, but this profile will only occur once in the unique set). The distribution of this index had substantial structure (as visible in Figure 1, and clear, particularly for plastids, in Figure S11), but we do not consider further transformations or more tailored analysis here, instead focusing on the similarity of results with those from the other index.

Prediction of retention index

We used Bayesian model selection with non-local priors to promote separation of overlapping models (Johnson and Rossell, 2010); specifically, moment (MOM) priors parameterised so that a signal-to-noise ratio of >0.2 is anticipated, compatible with previous findings (Johnston and Williams, 2016); a beta-binomial $(1,1)$ prior distribution on the model space, and a minimally informative inverse gamma prior for noise. Further information on priors, and the effects of varying them, are given in Figure S12. We implemented the selection process in the R package *mombf*.

In addition to the Bayesian linear model approach, we used a variety of different approaches for retention index regression. These included decision linear modelling with ridge and LASSO penalisation, decision tree regression, and random forest regression. The training, test, and cross-organelle performance of these approaches is given in Table S2.

Classification of subcellular encoding

We used Bayesian model averaging for generalised linear models (GLMs) predicting encoding compartments with priors giving probability $1/2$ for the inclusion of each parameter, implemented in *BMA*. We then trained GLMs involving hydrophobicity and carboxyl pK_a on a training subset of genes for each species. The training subset was the union of a random sample of half the nuclear-encoded genes and half the organelle-encoded genes in each species, with the test set being the complement of this set. We also used decision tree and random forest approaches for the classification task. For binding energy values, we used both a Bayesian GLM treating all complexes independently, with t-distributed priors with zero mean, implemented in *arm*; and a Bayesian generalised linear mixed model with flat priors over coefficients, residuals, and covariance structure, implemented in *blme*. These priors were used to overcome convergence issues given the perfect separation of datapoints observed for some protein complexes. Complexes were visualised in PyMOL (DeLano et al., 2002).

We also considered decision tree and random forest approaches for the organelle/nuclear encoding compartment classification problem; performance is shown in Table S3, with illustrations in Figure S7.

Endosymbionts and relatives

We considered a range of endosymbionts highlighted in a comprehensive recent review (Husnik and Keeling, 2019). For each we sought to identify a close free-living relative. In some cases all closest relatives of an endosymbiont themselves adopted a largely or obligate intracellular lifestyle; in these cases we tried to identify the closest relative that was at least capable of free-living (Table S4).

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantification and statistical analyses are an integral part of this research and each approach is described in detail above and as it arises. Briefly, we use a bioinformatic pipeline to curate oDNA datasets, constructing gene profiles for each species and collecting physical and chemical statistics for each gene. We create a quantitative retention index borrowing from hypercubic transition path sampling. We then use Bayesian model selection to (a) predict this index from those statistics using linear models and (b) predict encoding compartment from these statistics using generalised linear models. In this Bayesian paradigm, quantitative outputs are posterior probabilities over model indices and parameter values, and 95% credibility intervals are given; the effect of different priors is explored throughout. For binding energy analysis we quantify binding energy using PDBePISA and use both maximum likelihood

and Bayesian generalised linear models to relate the variables, with confidence and credibility intervals given accordingly. For the recent endosymbiont analysis we use Wilcoxon rank-sum tests to compare statistics in different compartments.

Code and dependencies

Code is written in R, Python, and C, with a wrapper script for bash, and is freely available at github.com/StochasticBiology/odna-loss, publication release https://doi.org/10.5281/zenodo.6873510.

Our pipeline uses the following R packages: ape (Paradis and Schliep, 2019), arm (Gelman and Su, 2020), blme (Chung et al., 2013), BMA (Raftery et al., 2021), caper (Orme et al., 2018), cowplot (Wilke, 2020), e1071 (Meyer et al., 2021), geiger (Pennell et al., 2014), GGally (Schloerke et al., 2021), ggnewscale (Campitelli, 2021), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2020), ggpval (Cheng, 2021), ggrepel (Slowikowski, 2021), ggtree (Yu et al., 2017), ggtreeExtra (Xu et al., 2021), glmnet (Friedman et al., 2010), gridExtra (Auguie, 2017), hexbin (Carr et al., 2021), igraph (Csardi and Nepusz, 2006), lme4 (Bates et al., 2015), logistf (Heinze et al., 2020), mombf (Rossell et al., 2021), nlme (Pinheiro et al., 2021), phangorn (Schliep, 2011), phytools (Revell, 2012), randomForest (Liaw and Wiener, 2002), stringdist (van der Loo, 2014), stringr (Wickham, 2019), and tree (Ripley, 2021).

We also use BioPython (Cock et al., 2009) for parsing sequences and computing gene statistics, PyMOL (DeLano et al., 2002) for visualisation, and BLAST (Camacho et al., 2009) for sequence comparisons.

Cell Systems, Volume *13*

Supplemental information

Evolutionary inference across eukaryotes identifies universal features shaping organelle gene retention

Konstantinos Giannakis, Samuel J. Arrowsmith, Luke Richards, Sara Gasparini, Joanna M. Chustecki, Ellen C. Røyrvik, and Iain G. Johnston

Supplementary Information

Method	MT training	MT test	PT training	PT test	PT predicting MT	MT predicting PT
LM (simple)	0.64	0.63	0.62	0.60	0.65	0.55
LM-pruned (simple)	0.73	0.71	0.72	0.72	0.68	0.50
LM (barcode)	0.71	0.69	0.58	0.56	0.72	0.59
LM-pruned (barcode)	0.71	0.70	0.64	0.64	0.67	0.51

Table S1: **Mean linear model regression performance, related to Fig. 2**. Spearman's ρ between predicted and observed retention index in test sets for different cases. Non-standard genes (*msh1/muts, matr, mttb*) are removed from mtDNA sets for these experiments. Labels show simple retention index vs barcode retention index; 'pruned' dataset (retaining only mt genes from families *nad, sdh, atp, cox, cytb, rp* and pt from *psa, psb, rp, rbc, ndh, atp, pet*) vs unpruned. Each LM uses only GC content and hydrophobicity.

Method	MT training	MT test	PT training	PT test	PT predicting MT	MT predicting PT
Tree	0.79	0.40	0.82	0.45	0.54	0.33
LM	0.70	0.43	0.71	0.66	0.52	0.25
Tree-reduced	0.73	0.48	0.75	0.45	0.55	0.39
LM-Reduced	0.58	0.52	0.61	0.61	0.54	0.48
Ridge	0.68	0.39	0.66	0.71	0.57	0.41
LASSO	0.63	0.44	0.66	0.71	0.57	0.37
SVR	0.81	0.46	0.77	0.62	0.62	0.34
RF	0.92	0.48	0.95	0.62	0.62	0.45
RF-Reduced	0.88	0.50	0.92	0.51	0.57	0.50
RF-Cross	0.94	N/A	0.96	N/A	0.62	0.56
RF-Cross-Reduced	0.90	N/A	0.92	N/A	0.55	0.59

Table S2: **Mean regression performance (Spearman's ρ between predicted and observed indices) predicting retention index with different approaches, related to Fig. 2.** Non-standard genes (*msh1/muts, matr, mttb*) are not removed for these experiments. Tree, decision tree regression; LM, linear model; Ridge, ridge regression; LASSO, LASSO regression; RF, random forest regression. All genetic features included by default; 'reduced' corresponds to models involving only GC content and hydrophobicity. 'Cross' refers to cross-organelle experiments where mt training is used to predict pt test and vice versa (N/A, not applicable: no test set within training organelle).

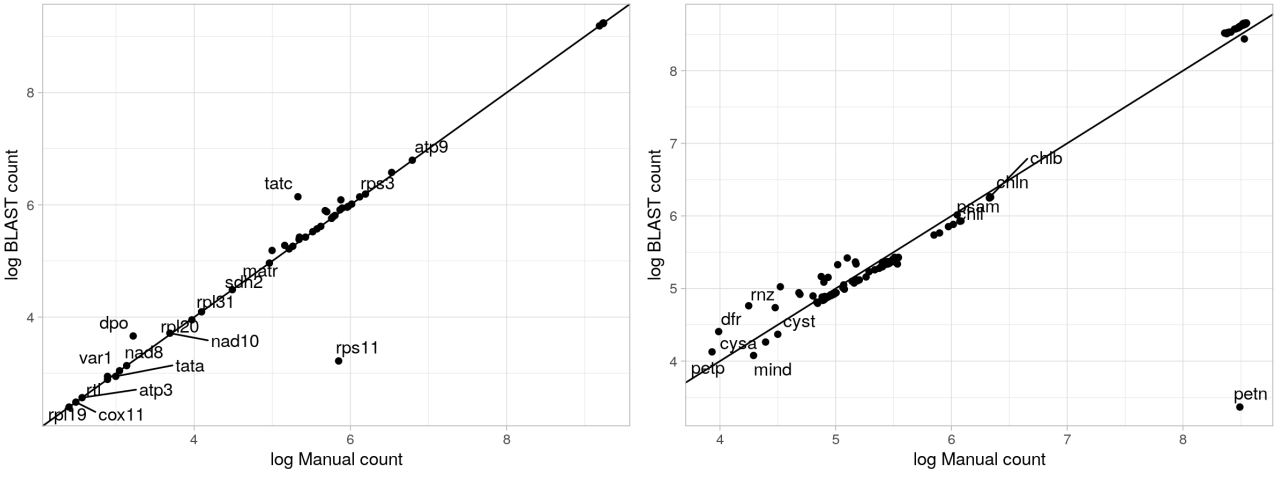


Figure S1: **Correlation between gene counts across species derived using manual and BLAST labelling approaches, related to Fig. 1**. r = 0.9999 for mitochondrial and r = 0.9849 for plastid data; discrepancies are largely down to a small number of outliers.

Complex	Model type	Training	Test	Balance	Complex	Model type	Training	Test	Balance
nad[0-9]	tree	0.99	0.99	0.10	nad[0-9]	RF	1.00	1.00	0.10
sdh[0-9]	tree	0.97	0.91	0.66	sdh[0-9]	RF	1.00	0.95	0.68
cytb	tree	0.99	0.99	0.18	cytb	RF	1.00	0.99	0.18
cox[0-9]	tree	1.00	0.99	0.09	cox[0-9]	RF	1.00	0.99	0.09
atp[0-9]	tree	0.98	0.96	0.16	atp[0-9]	RF	1.00	0.98	0.16
(MT) rp[sl]	tree	0.88	0.85	0.69	(MT) rp[sl]	RF	1.00	0.92	0.69
psa[a-x]	tree	0.99	0.99	0.03	psa[a-x]	RF	1.00	0.99	0.03
psb[a-z]	tree	1.00	0.99	0.01	psb[a-z]	RF	1.00	1.00	0.01
atp[a-z]	tree	0.98	0.97	0.12	atp[a-z]	RF	1.00	0.99	0.12
pet[a-z]	tree	1.00	0.99	0.01	pet[a-z]	RF	1.00	0.99	0.01
rbc	tree	0.99	0.97	0.07	rbc	RF	1.00	0.98	0.07
(PT) rp[sl]	tree	0.99	0.99	0.02	(PT) rp[sl]	RF	1.00	0.99	0.02
nad[0-9]	tree-reduced	0.99	0.99	0.10	nad[0-9]	RF-reduced	1.00	0.99	0.10
sdh[0-9]	tree-reduced	0.97	0.92	0.66	sdh[0-9]	RF-reduced	1.00	0.93	0.66
cytb	tree-reduced	0.98	0.97	0.18	cytb	RF-reduced	1.00	0.98	0.19
cox[0-9]	tree-reduced	0.98	0.98	0.09	cox[0-9]	RF-reduced	1.00	0.98	0.09
atp[0-9]	tree-reduced	0.92	0.91	0.16	atp[0-9]	RF-reduced	1.00	0.92	0.16
(MT) rp[sl]	tree-reduced	0.79	0.76	0.69	(MT) rp[sl]	RF-reduced	1.00	0.77	0.69
psa[a-x]	tree-reduced	0.98	0.97	0.03	psa[a-x]	RF-reduced	1.00	0.97	0.03
psb[a-z]	tree-reduced	0.99	0.99	0.01	psb[a-z]	RF-reduced	1.00	0.99	0.01
atp[a-z]	tree-reduced	0.91	0.90	0.12	atp[a-z]	RF-reduced	1.00	0.91	0.12
pet[a-z]	tree-reduced	0.99	0.99	0.01	pet[a-z]	RF-reduced	1.00	0.99	0.01
rbc	tree-reduced	0.96	0.93	0.06	rbc	RF-reduced	1.00	0.94	0.07
(PT) rp[sl]	tree-reduced	0.98	0.98	0.02	(PT) rp[sl]	RF-reduced	1.00	0.98	0.02
All PT	tree-cross	0.94	0.80	N/A	All PT	RF-cross	1.00	0.60	N/A
All MT	tree-cross	0.98	0.82	N/A	All MT	RF-cross	1.00	0.79	N/A
All PT	tree-cross-reduced	0.94	0.56	N/A	All PT	RF-cross-reduced	1.00	0.47	N/A
All MT	tree-cross-reduced	0.97	0.81	N/A	All MT	RF-cross-reduced	1.00	0.82	N/A

Table S3: **Nuclear-organelle classification performance, related to Fig. 3.** Proportion of test set assigned to correct compartment, by organelle complex, with different approaches (tree, decision tree; RF, random forest). Complexes are labelled with regular expressions describing their gene labels. All genetic features included by default; 'reduced' corresponds to models involving only GC content and hydrophobicity. 'Cross' refers to cross-organelle experiments where mt training is used to predict pt test and vice versa. Balance gives the proportion of genes encoded in one compartment (may fluctuate slightly due to different subsamples being used in model construction): N/A, not applied to cross-organelle classification.

Endosymbiont	NCBI accession	Free-living relative	NCBI accession	References
Nasuia deltocephalinicola	CP013211.1	*Herbaspirillum seropedicae*	CP002039.1	(Bennett and Moran, 2013)
Ca. Sulcia muelleri	CP001981.1	*Porphyromonas gingivalis*[1]	AE015924.1	(McCutcheon and Moran, 2007)
Ca. Tremblaya phenacola	CP003982.1	*Sodalis praecaptivus*	CP006569.1	(Enomoto et al., 2017)
Rhopalodia gibberula SB	AP018341.1	*Cyanothece sp. PCC 8801*	CP001287.1	(Nakayama and Inagaki, 2017)
Ca. Hodgkinia cicadicola	CP008699	*Rhizobium etli*	CP007641.1	(McCutcheon et al., 2009)
Ca. Pinguicoccus supinus	CP039370.1	*Coraliomargarita akajimensis*[2]	CP001998.1	(Serra et al., 2020)
Ca. Fokinia solitaria	CP025989.1	*Pelagibacter ubique*[3]	CP000084.1	(Floriano et al., 2018)
Paulinella chromatophore	CP000815.1	*Synechococcus PCC 7002*	CP000951	(Lhee et al., 2019)
Ca. Azoamicus ciliaticola	NZ_LR794158.1	*Legionella clemsonensis*[4]	NZ_CP016397	(Graf et al., 2021)
Nostoc azollae	CP002059.1	*Raphidiopsis brookii*	ACYB01000001.1	(Ran et al., 2010)

Table S4: **Independent endosymbionts and close free-living relatives, related to Fig. 4.** SB, spherical body. [1] Relative does invade cells but can survive in oral cavity. [2] Partner is not closest sequence found, but is closest annotated sequence in putative phylogeny. [3] All closest relatives are intracellular Rickettsiales – relative taken from a sister group. [4] Most relatives, including Legionella, are largely intracellular.

		Hydro	Hydro_I	Molecular weight / Da	pKa1	pKa2	Aglucose	CWEnergy
Ala	A	41	3	89.1	2.34	9.69	0.5	12.5
Arg	R	-14	1	174.2	2.17	9.04	1.39	18.5
Asn	N	-28	1	132.12	2.02	8.8	0.79	4
Asp	D	-55	1	133.11	1.88	9.6	0.61	1
Cys	C	49	3	121.16	1.96	10.28	0.75	24.5
Gln	Q	-10	2	146.15	2.17	9.13	0.92	9.5
Glu	E	-31	1	147.13	2.19	9.67	0.86	8.5
Gly	G	0	2	75.07	2.34	9.6	0.31	14.5
His	H	8	2	155.16	1.82	9.17	1.46	33
Ile	I	99	4	131.18	2.36	9.6	1.21	20
Leu	L	97	4	131.18	2.36	9.6	1.21	33
Lys	K	-23	1	146.19	2.18	8.95	1.31	18.5
Met	M	74	4	149.21	2.28	9.21	1.25	18.5
Phe	F	100	4	165.19	1.83	9.13	1.84	63
Pro	P	-46	1	115.13	1.99	10.6	0.99	12.5
Ser	S	-5	2	105.09	2.21	9.15	0.49	15
Stop	X	-	-	-	-	-	-	-
Thr	T	13	2	119.12	2.09	9.1	0.69	6
Trp	W	97	4	204.23	2.83	9.39	2.39	78.5
Tyr	Y	63	3	181.19	2.2	9.11	1.77	56.5
Val	V	76	4	117.15	2.32	9.62	0.96	25

Table S5: **Amino acid properties used in model selection, related to Figs. 1-3.** Numerical values of the properties described in the text. Quantities are unitless unless specific. See text for sources. Feature codes are [Hydro]phobicity, [Hydro_I] hydrophobicity index, [pKa1] carboxyl pKa, [pKa2] amino pKa, [MW] molecular weight, [Aglucose, CWEnergy] energies of gene expression (Barton et al., 2010; Craig and Weber, 1998).

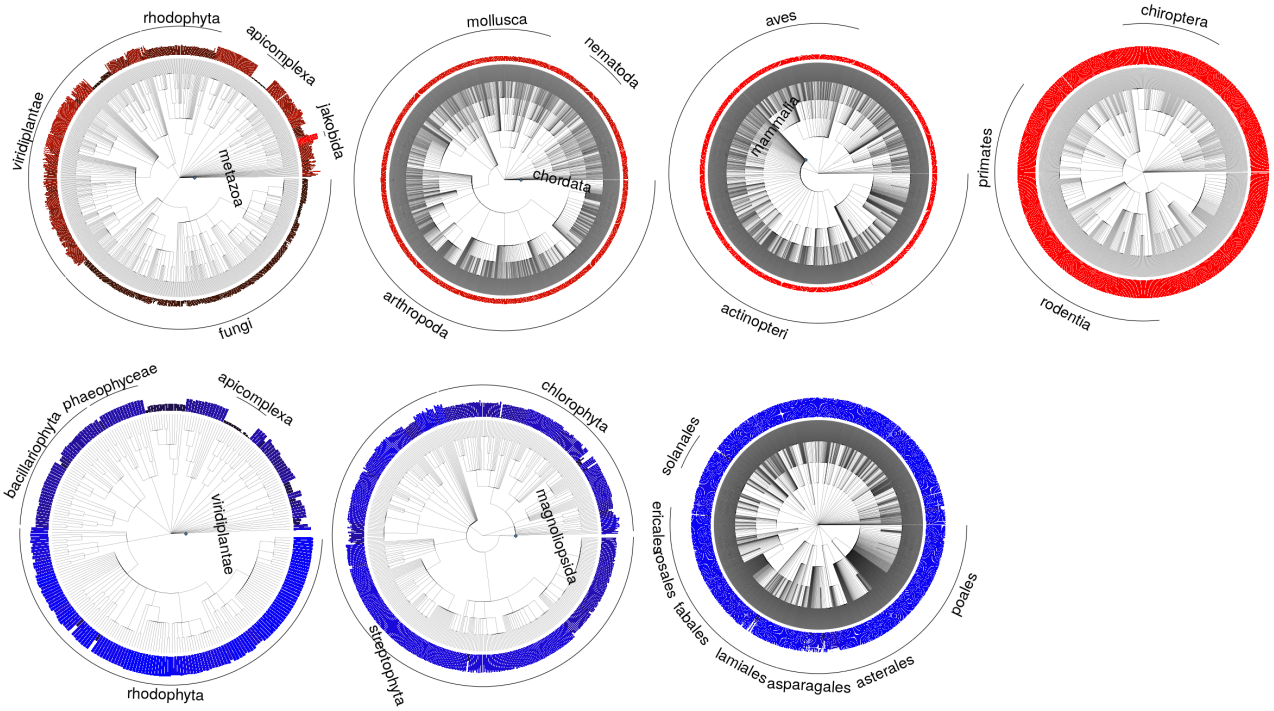


Figure S2: **Taxonomic trees for the mt and pt datasets, related to Fig. 1**. Blue diamonds give truncation points; associated taxa are expanded in the next rightward tree. Truncated taxa are broadly chosen to reflect those with less diversity in oDNA. Bars illustrate number of retained organelle genes in each species (scale differs in each subtree).

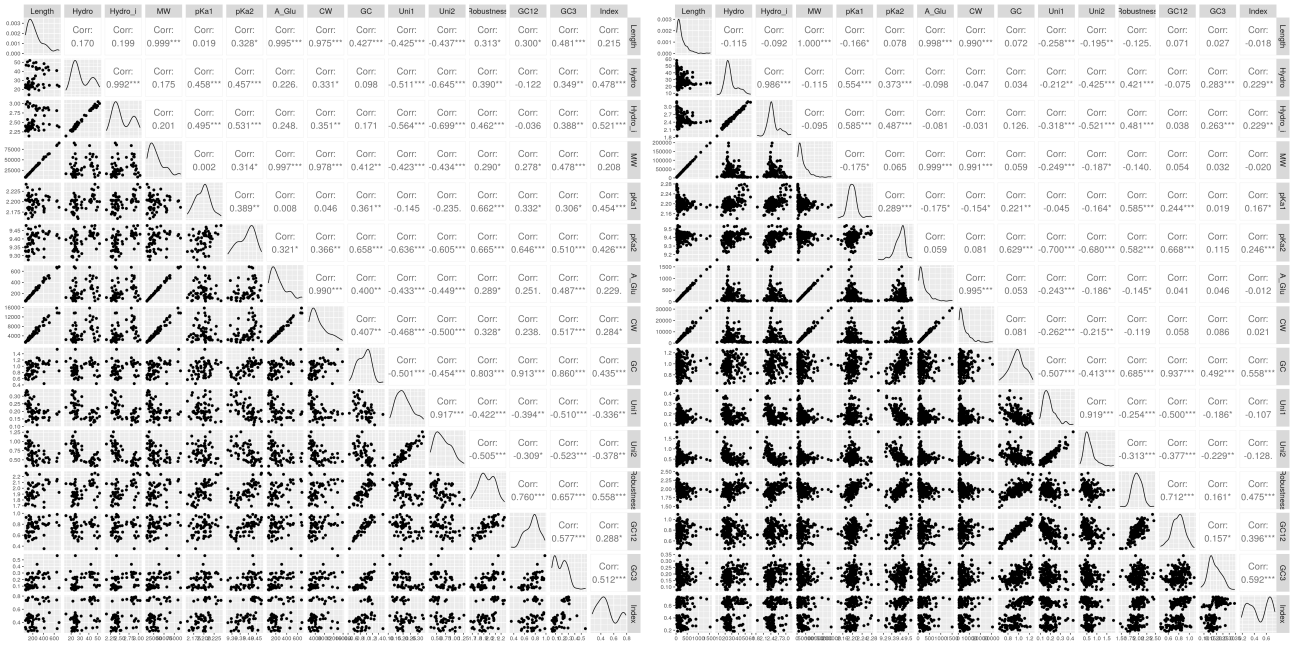


Figure S3: **Linear correlations between genetic features and retention index, for mt and pt genes, related to Fig. 2.**

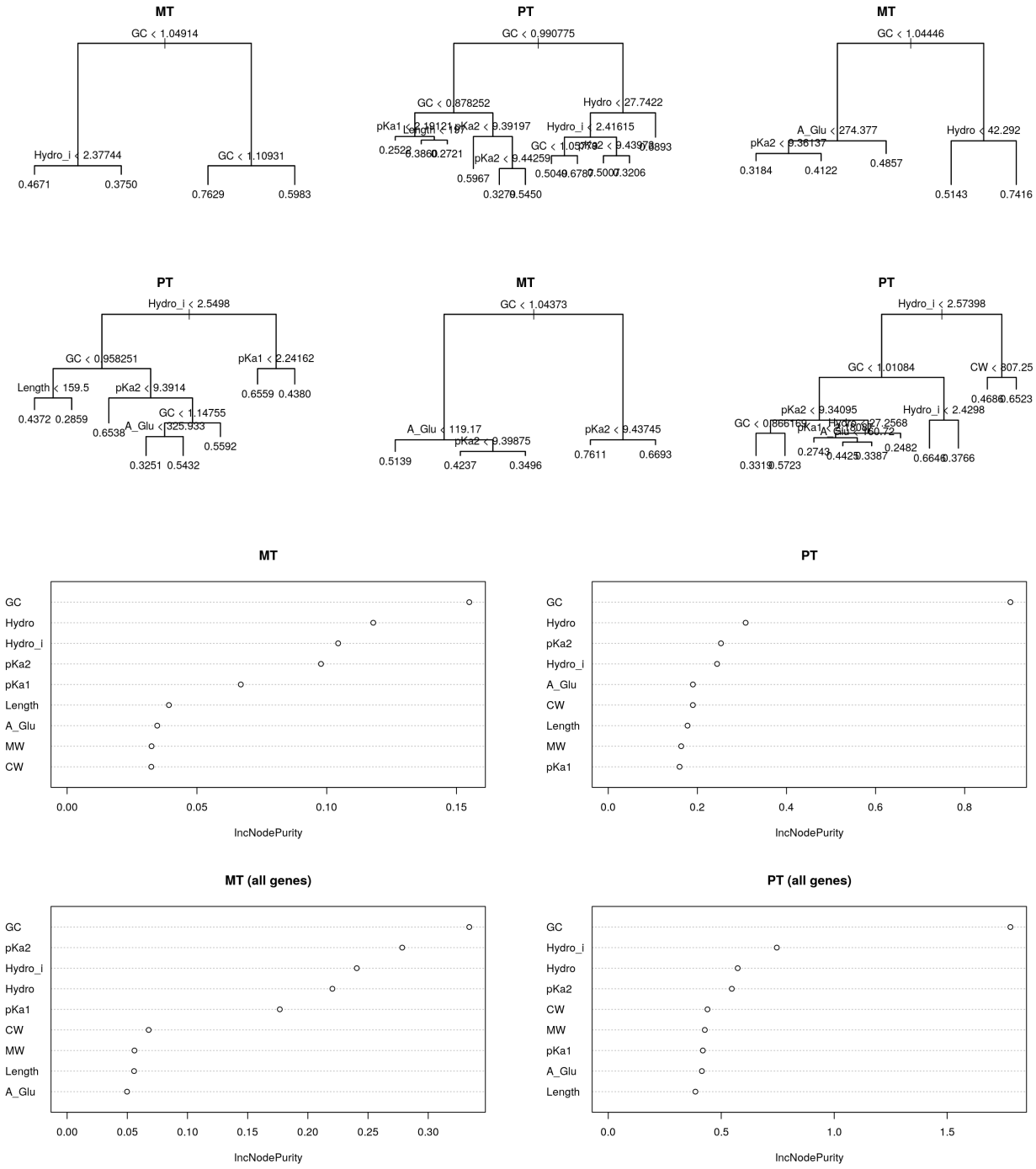


Figure S4: **Decision tree and random forest regression for retention index, related to Fig. 2.** (top) a set of trees learned to predict retention for different training-test splits, showing the dominant role of GC content and hydrophobicity as predictive features. (bottom) variance improvement plots for random forest regression of the same task, illustrating the importance of each feature in the predictive outcome.

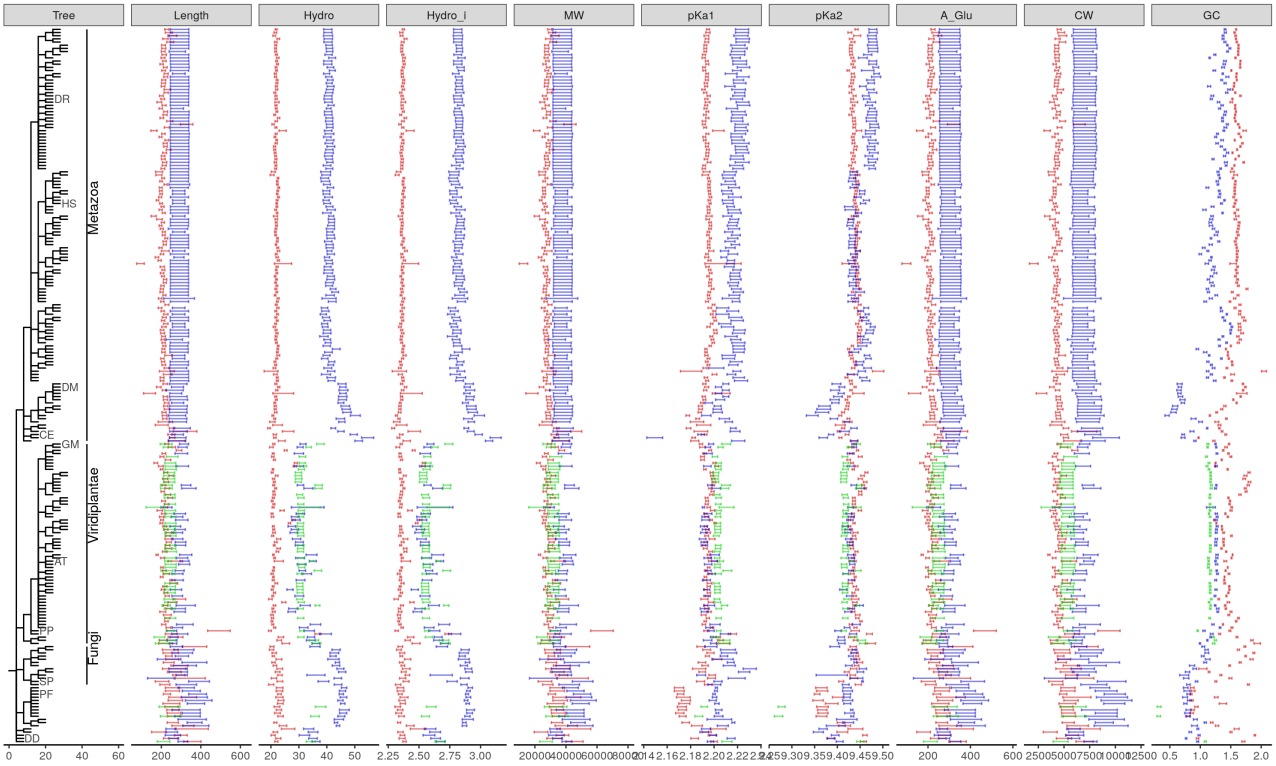


Figure S5: **Statistics of genes encoded in the nucleus (red), mitochondrion (blue), or plastid (green) compartments, related to Fig. 3.** Bars give mean and s.e.m. for each species; phylogeny shows the relationship between species. Specific model species labelled by initials: *Danio rerio*, *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Glycine max*, *Arabidopsis thaliana*, *Physcomitrella patens*, *Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Dictyostelium discoideum*.

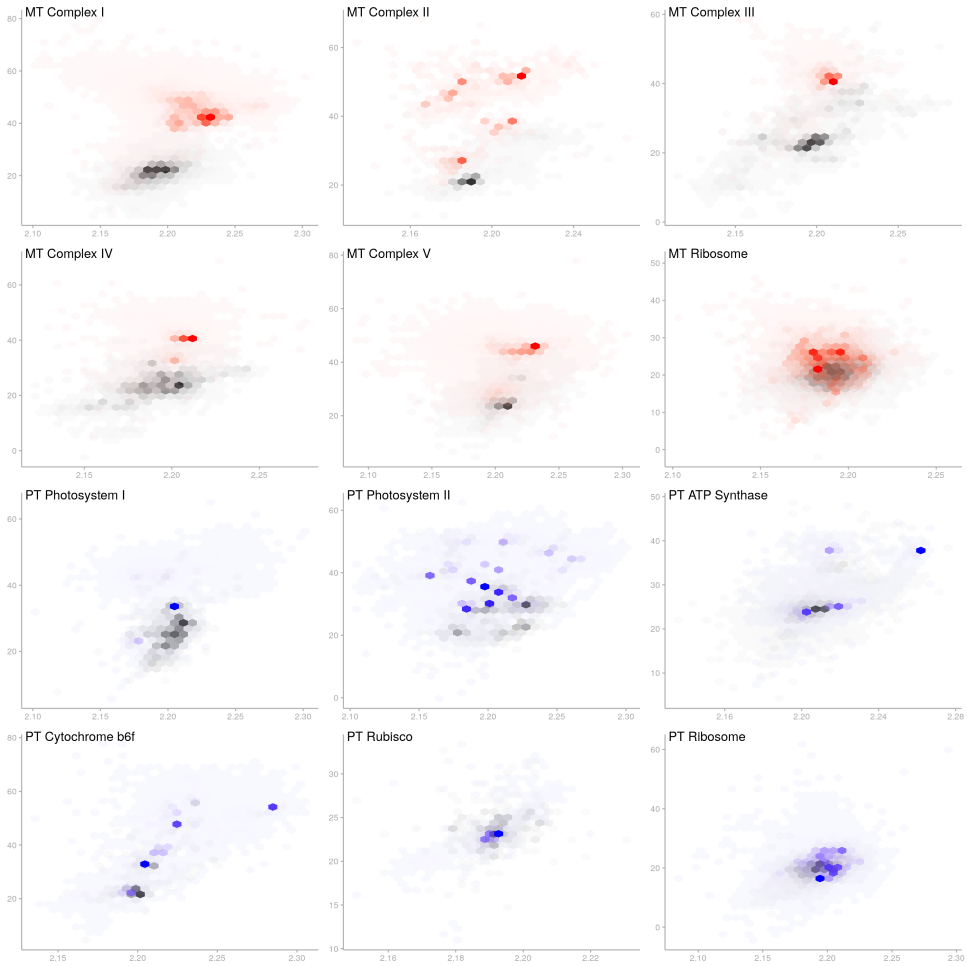


Figure S6: **Hydrophobicity and carboxyl pKa for nuclear- and organelle-encoded complex subunits, related to Fig. 3.**

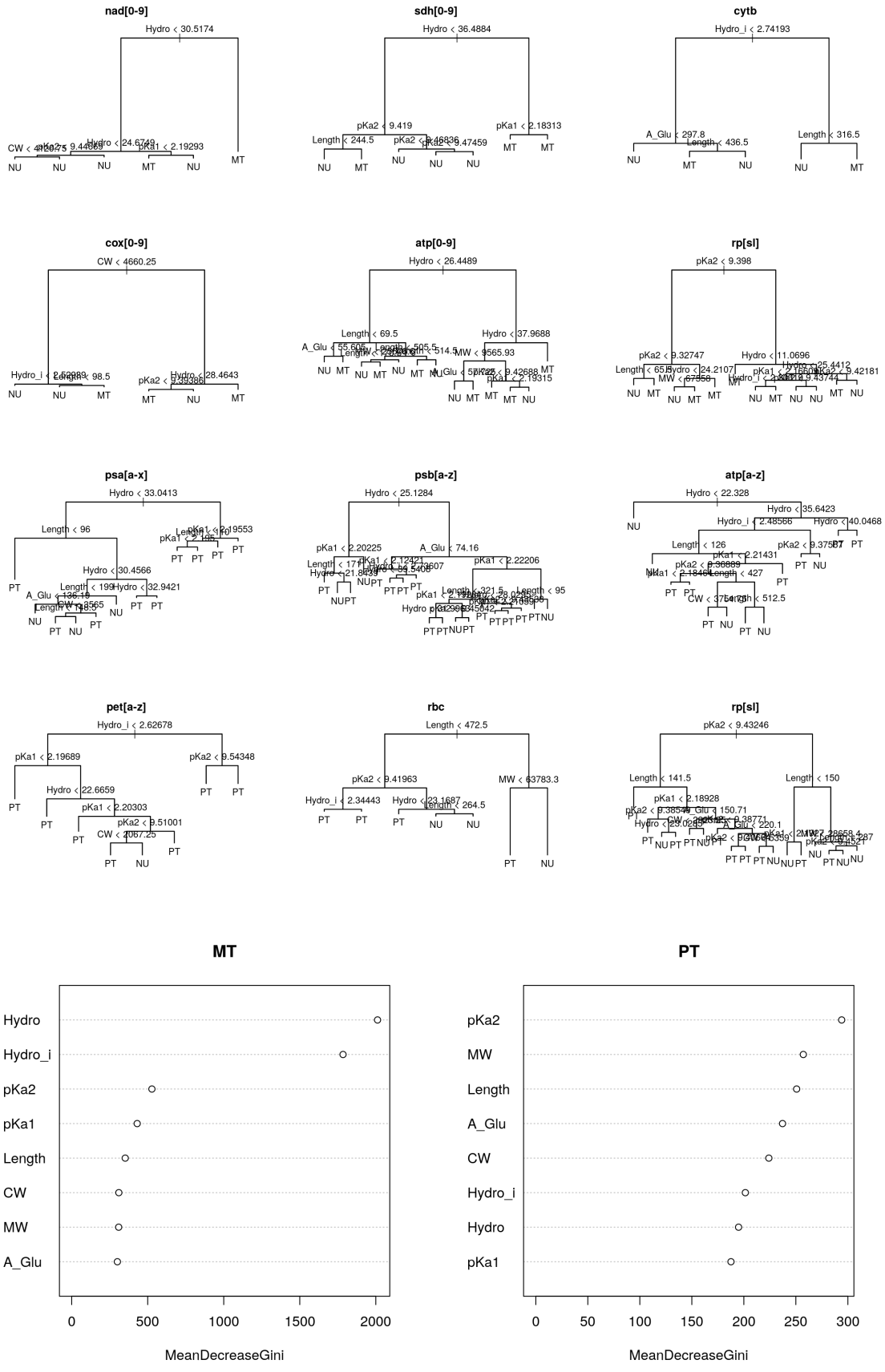


Figure S7: **Decision tree and random forest classification for encoding compartment, related to Fig. 3.** (top) a set of trees learned to predict encoding compartment for genes in different protein complexes, showing roles for hydrophobicity, pKa, and production energy (CW) as predictive features. (bottom) variance improvement plots for random forest regression for compartment classification across all genes, illustrating the importance of each feature in the predictive outcome. Complexes are labelled with regular expressions describing their gene labels.

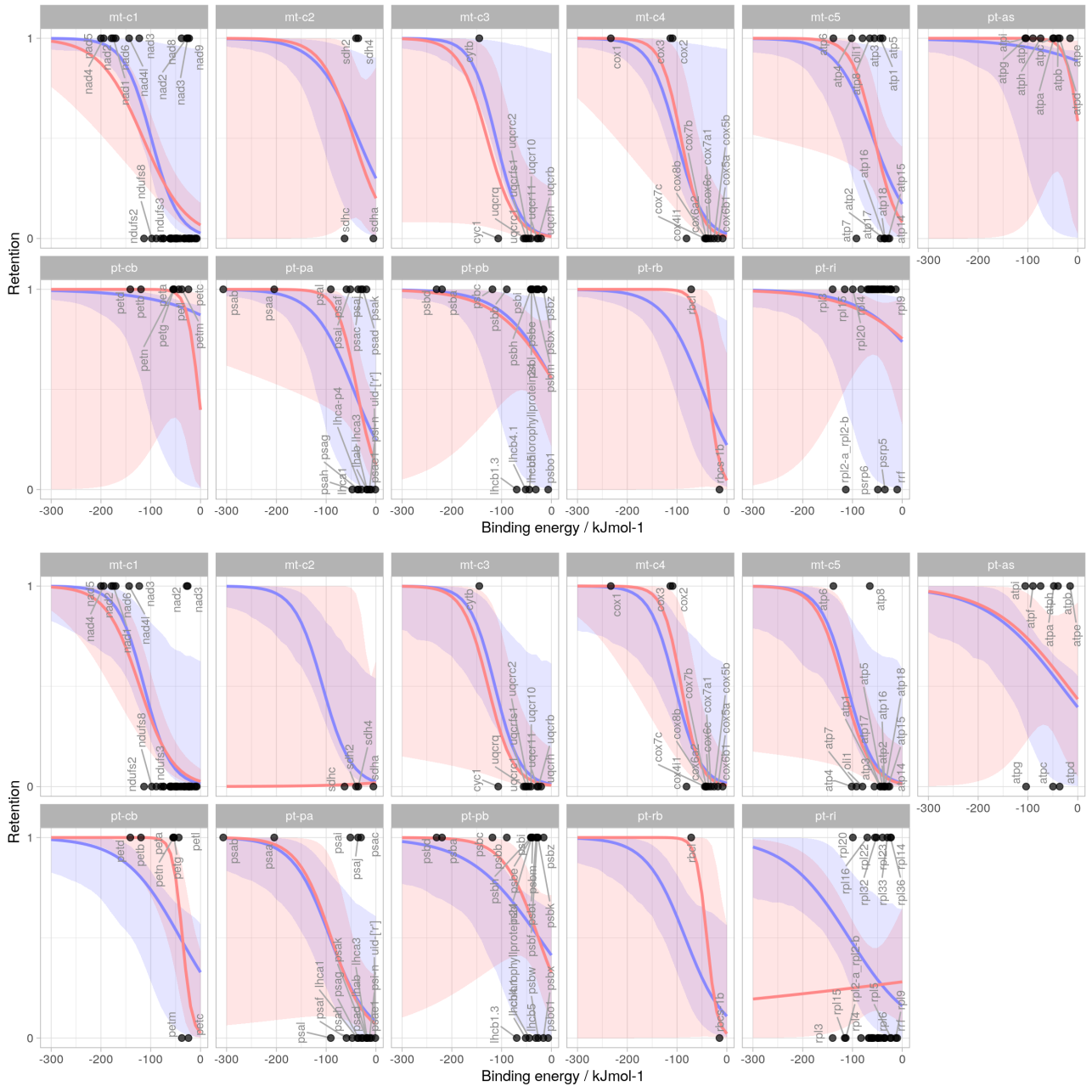


Figure S8: **Binding energy links to encoding compartment, related to Fig. 3.** (top) Comparison of Bayesian generalised linear model (GLM) and generalised linear mixed model (GLMM) for binding energy-retention relationship. The GLM approach (red) treats each complex independently; the GLMM (blue) describes complex-specific changes to an overall trend. Frequentist p-values against the null hypothesis of no relationship are 0.00047 (GLM) and 0.0038 (GLMM). (bottom) The same plot, but using a threshold of 1000 species rather than 1 species in our dataset retaining a gene for it to be assigned a retention value of 1. The corresponding p-values are 3.1×10^{-5} (GLM) and 0.0023 (GLMM).

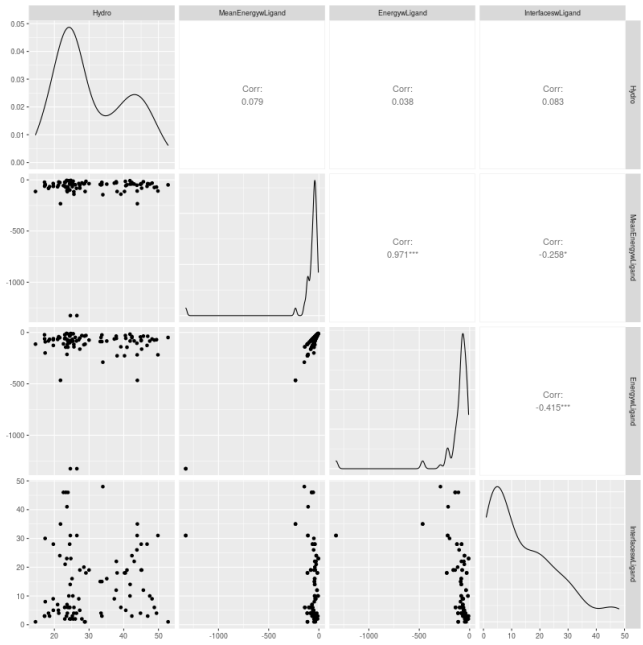


Figure S9: **Little correlation between hydrophobicity and energetic centrality across gene products involved in the complexes studied, related to Fig. 3.**

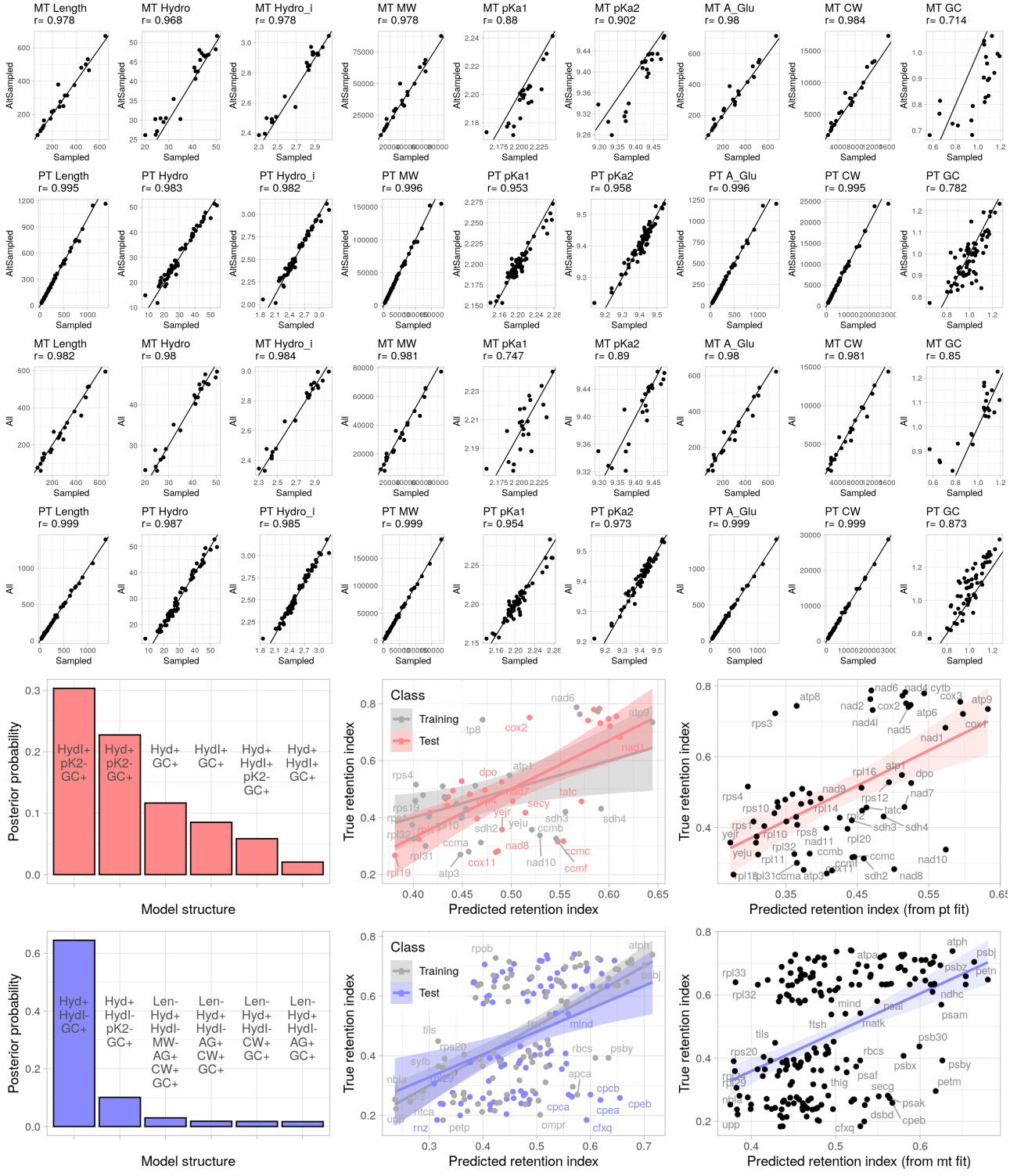


Figure S10: **Effect of different averaging protocols to summarise gene statistics, related to Fig. 2.** (top) Correlations between our model systems average and (A) average across randomly sampled species from different clades and (B) average across all species in the dataset (expected to be highly weighted towards bilaterians and angiosperms). (bottom) Result of model selection and testing process with the random species averaging protocol.

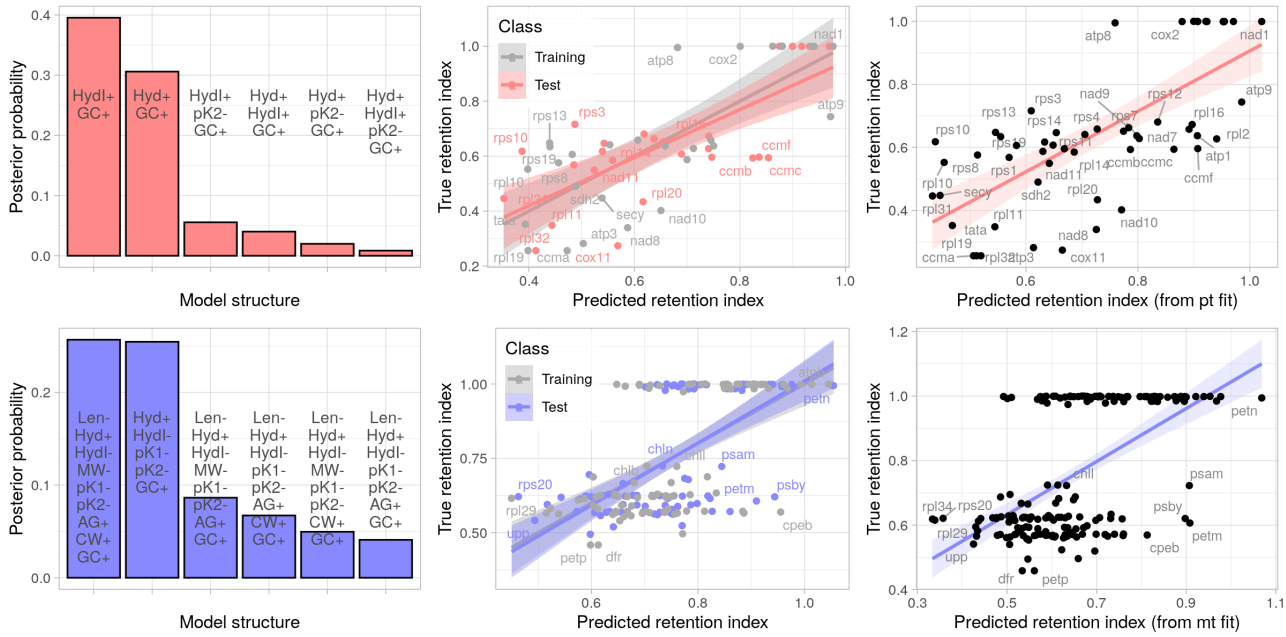


Figure S11: **Model selection and regression for the barcode-based retention index reflects the outcomes from the simple retention index, related to Fig. 2.** See also Table S1.

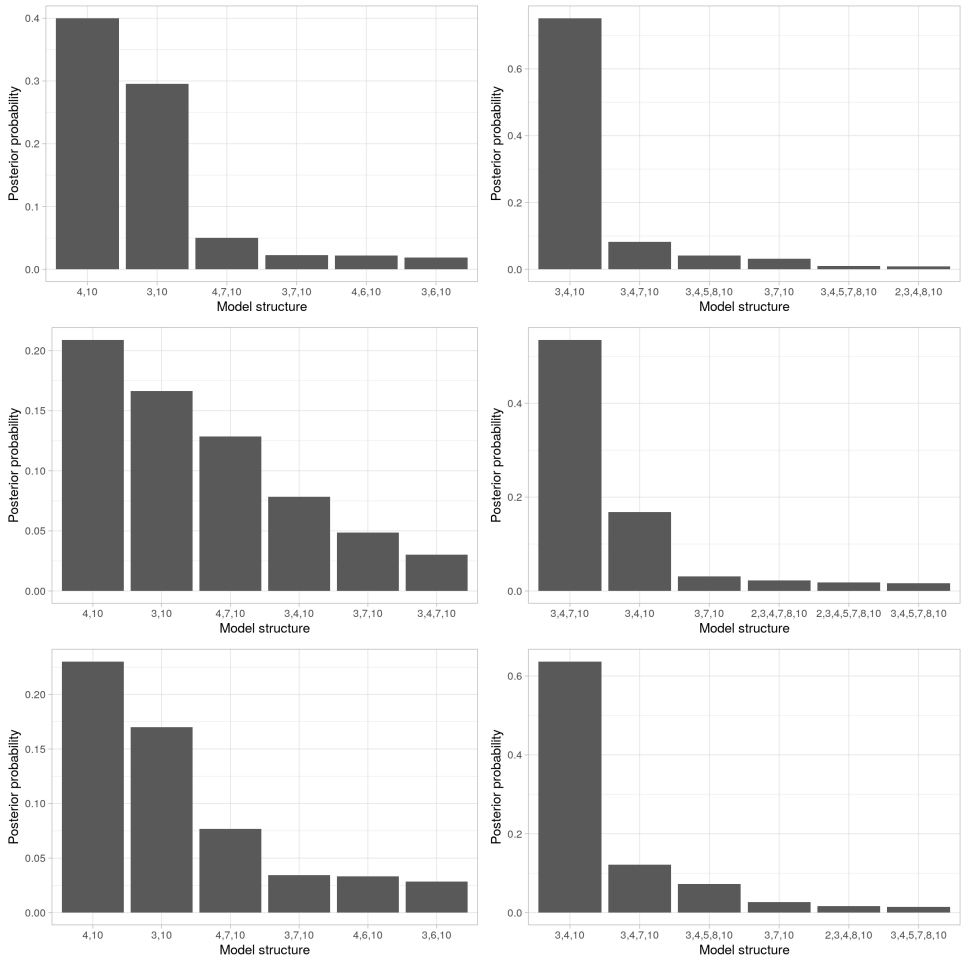


Figure S12: **Bayesian model selection for linear models predicting retention index, with different priors from the default choices in the main text, related to Fig. 2.** Left, mitochondrial; right, plastid data. Top, inverse moment (iMOM) prior with τ = 0.133 and beta-binomial(1,1) prior over models. Centre, moment (MOM) prior with τ = 0.348 and uniform prior over models. Bottom, IMOM prior with τ = 0.133 and uniform prior over models. MOM vs iMOM changes structure of non-local priors; model priors assign different prior weights to overall model structures. Features appearing in models are: 1 (intercept); 2 (length); 3 (hydrophobicity); 4 (hydrophobicity index); 5 (molecular weight); 6 (amino pKa); 7 (carboxyl pKa); 8 (glucose assembly energy); 9 (alternate assembly energy); 10 (GC content).