bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Evolutionary inference across eukaryotes identifies specific pressures favoring mtDNA gene retention

Iain G. Johnston[1,*], Ben P. Williams[2]

[1] School of Biosciences, University of Birmingham, United Kingdom

[2] Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, United States of America

[*] Correspondence to i.johnston.1@bham.ac.uk

Abstract

Since their endosymbiotic origin, mitochondria have lost most of their genes. Despite the central cellular importance of mtDNA, selective mechanisms underlying the evolution of its content across eukaryotes remain contentious, with no data-driven exploration of different hypotheses. We developed HyperTraPS, a powerful methodology coupling stochastic modelling with Bayesian inference to identify the ordering and causes of evolutionary events. Using a large dataset from over 2000 complete mitochondrial genomes, we inferred evolutionary trajectories of mtDNA gene loss across the eukaryotic tree of life. We find that proteins fulfilling central energetic roles in the assembly of mitochondrial complexes are prefentially retained in mitochondria across eukaryotes, biophysically supporting 'colocalization for redox regulation'. We also provide the first quantitative evidence that a combination of GC content and protein hydrophobicity is required to explain mtDNA gene retention, and predicts the success of artificial gene transfer experiments. Our results demonstrate that a combination of these three characteristics explains most mtDNA gene variability, addressing the long-debated question of why particular genes are retained in mitochondrial genomes.

Introduction

Mitochondria are the result of an endosymbiotic event [1], where a free-living organism resembling an α-proteobacterium was engulfed by another cell billions of years ago. Since this event, a pronounced loss of mitochondrially-encoded genes has occurred, with genes either transferred to the host nucleus or lost completely [2, 3, 4, 5]. This endosymbiosis and the subsequent evolution of mitochondrial genomes are among the most important processes in biological history, giving rise to eukaryotic life and hypothesised as facilitating the higher energy output required for the evolution of complexity and multicellularity [6].

The precursor mitochondrion is estimated to have possessed thousands of genes [7]. Present-day mitochondrial genomes retain only a small and highly variable number of these: malarial parasites possess only three protein-coding genes in their mtDNA [8] and some protists possess over sixty [9] (human mtDNA encodes thirteen protein-coding genes [10]). Some eukaryotes completely lack mtDNA, representing the limiting case in our picture of mitochondrial gene loss; there is intriguing evidence that changes in nuclear DNA can compensate for a lack of mtDNA in some of these cases [11]. Despite the importance of mitochondrial evolution in diverse fields including phylogenetics [12] and human medicine [10], the forces influencing mtDNA content are still debated and poorly understood, limiting our understanding of the evolutionary history of eukaryotic bioenergetics. Dual questions exist, both as to why so many genes have been lost from the mitochondrion, and why some genes are retained in mtDNA [13].

The first question, why genes are lost from mitochondrial genomes, has a set of answers that are largely agreed upon [14], and several plausible mechanisms [15, 16, 17]. Selective advantages to the nuclear encoding of organellar genes include the avoidance of Muller's ratchet (the irreversible buildup of deleterious mutations) [12, 13], protection from mitochondrial mutagens [18], and enhanced fixing of beneficial mutations [14, 13].

Answers to the second question, why some genes are retained in organelles, remain more elusive. The simplest possibility, which is sometimes assumed, is the 'null hypothesis' that gene loss is uniform and random, with no particular link between gene properties and retention in mtDNA. Many possible alternative hypotheses are currently discussed, with two possibilities in particular finding the most qualitative support [19]. The first proposes that highly hydrophobic proteins, if produced remotely, are difficult to import and sort across membranes [14, 20], or readily mistargeted to other organelles [21, 22], favouring the organellar retention of genes encoding these hydrophobic proteins, though the hypothesis is debated [23]. The second hypothesis is known as 'colocalisation for redox regulation' (CoRR [23]), suggesting that the retention of key organellar genes allows beneficial localised control of energetic machinery, so that the performance of individual organelles can be optimised with-

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

out affecting the whole-cell population. Recent work has supported this hypothesis, showing that genes encoding subunits central to the assembly of the mitochondrial ribosome are most retained [24]. Other hypotheses include the possible toxicity of some gene products in the cytosol [25], and differences between the genetic code of the mitochondrion and the nucleus leading to difficulties in interpretation [14]. Additional constraints on mitochondrial gene evolution have been identified including patterns in GC skew, suggested to result from asymmetric mutation pressure [26, 12], GC content, suggested to influence the free energy and thus stability and mutational susceptibility of mtDNA [27], and gene expression, suggested to modulate selective pressure on individual gene sequences in animals [28].

Quantitative evidence supporting one or more of these hypotheses over others is currently absent, with (highly debated) qualitative discussion constituting most of the current field. Existing studies have analysed specific gene loss events [29], proposed bioinformatic approaches for the analysis of sequence-level features in organellar genomes [30, 31], and explored purifying selection at the mtDNA nucleotide level in animals [28], and, in a broad-ranging study, explored the phylogenetic history and structure of mtDNA using gene clusters and intron structure [32]. However, we are unaware of any existing quantitative approach that identifies the evolutionary pressures behind the highly variable patterns of presence and absence of mtDNA genes across the whole of the eukaryotic tree of life. Although important progress has been made at a sequence level (Ref. [33] takes a phylogenetic approach in studying the sequence evolution of mitochondrial ribosomes across a diverse range of taxa), molecular phylogenies are not necessarily useful tools to answer broader questions about the evolution of genome structure [34], particularly given the broad range of taxa and possible parallel evolutionary pathways involved [35, 24]. To make progress with this complex evolutionary system, we developed highly generalisable stochastic and statistical machinery for inference of evolutionary dynamics, and applied it to a large dataset of over 2000 sequenced mtDNA genomes across eukaryotic life, reconstructing the evolutionary history of mitochondrial genomes. We identified genetic features corresponding to existing and novel hypotheses on the pressures driving mtDNA evolution, and used Bayesian model selection to identify the features, and thus the corresponding pressures, most likely to give rise to the inferred dynamics. We will show that a combination of GC content, hydrophobicity, and energetic centrality (each of which provides independent explanatory power) accounts for most of the variability in observed mtDNA structure and predicts the success of artificial gene transfer experiments.

Results

HyperTraPS: an algorithm for sampling rare evolutionary paths on a hypercubic transition network

To explore the evolutionary processes that give rise to present-day patterns of mtDNA genes across taxa, we constructed a general model for mtDNA evolution that allowed us to amalgamate and unify the large volume of genomic data available (see Methods). This model includes a description of every possible pattern of presence and absence of all 65 protein-coding mtDNA genes that we consider. These patterns are represented by binary strings, where a 0 as the ith character corresponds to the ith gene being absent, and a 1 at that position corresponds to that gene being present. We will use this picture to represent the transitions that make up the evolutionary history of mtDNA (Fig. 1 A). We allow evolutionary transitions between states that differ by one trait, so that individual genes are lost one-by-one: however, simultaneous loss of several genes can be captured by this model, and corresponds to an equal probability that each gene is lost at a given timestep. Each transition has a given *intensity*, and the probability of a given transition from some state is proportional to that transition's intensity (normalised by the sum of intensities of all transitions leading away from that state). Evolution is modelled as a walk from the state of all 1s to the state of all 0s, with individual transitions along the walk occurring randomly, weighted by intensity.

For example, consider a simple system restricted to $L = 3$ genes. The evolutionary space consists of eight states and ten possible transitions, illustrated in Fig. 1 B. If the $111 \to 110$ transition has intensity 0.98, and the $111 \to 101$ and $111 \to 011$ transitions each have intensity 0.01 (Fig. 1 Bi), we expect 98% of evolutionary processes to follow the initial step $111 \to 110$. Subsequent steps from 110 will likewise occur with probability proportional to their relative intensities, as illustrated in Fig. 1 B: in Fig. 1 Bi a step from 110 is highly likely to be $110 \to 100$ due to that transition's high intensity, whereas in Fig. 1 Bii (with a different set of intensities) a step from 011 is equally likely to be $011 \to 010$ or $011 \to 001$, as these transitions have equal intensity. The reader will note that the graphical representations of this model in Fig. 1 B have a cubic structure: as L increases this structure expands to become an L-dimensional hypercube.

The goal of this section is to construct an algorithm that allows us to compute how likely an evolutionary path within this modelling framework is to visit two given states. For example, consider the case where in our $L = 3$ example above strings describe the presence or absence of *geneA*, *geneB* and *geneC* in that order. We may have found that an ancestor possessed *geneA*, *geneB*, and *geneC* (111), and a descendant possesses *A* and *geneB* (110). It is trivial in this case to see that,

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

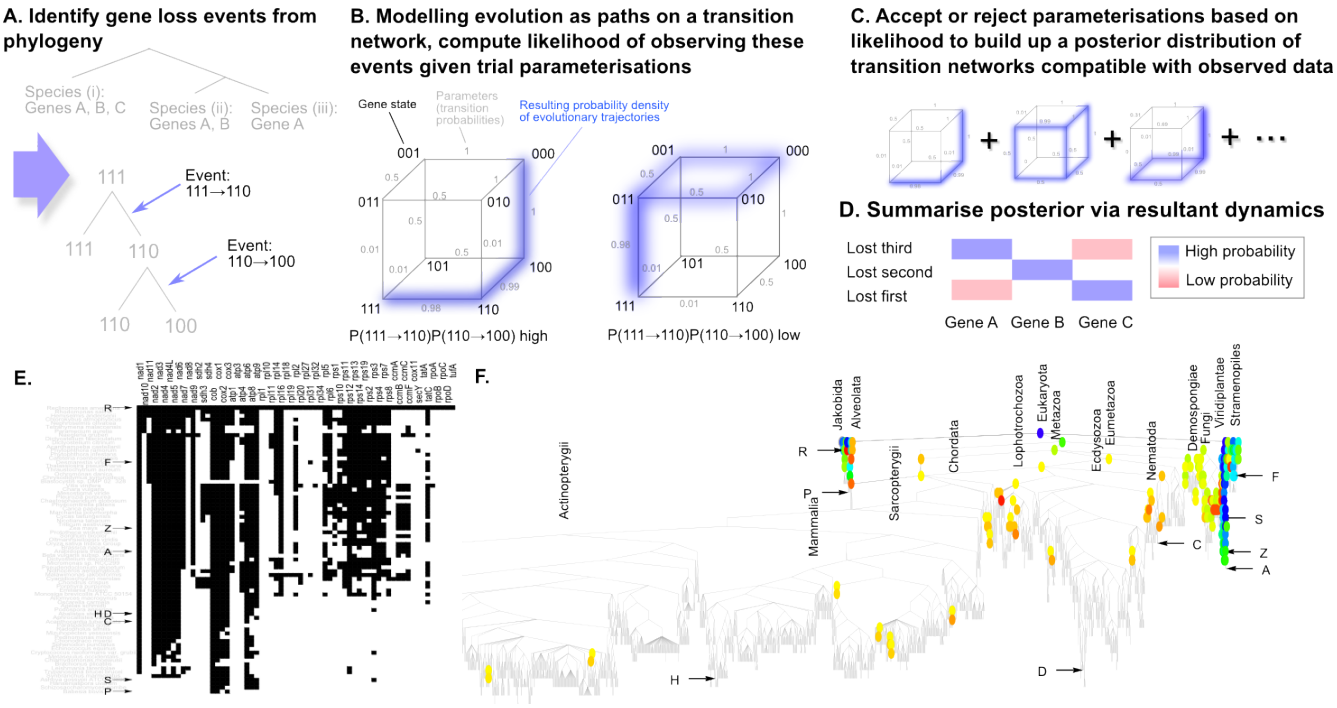

Figure 1: **Illustration and source data for HyperTraPS inference of mitochondrial gene loss ordering.** (A) Gene loss events are identified by inferring ancestral states on a given phylogeny, providing a set of observed transitions between gene states. (B) An evolutionary space defined by the presence or absence of $L = 3$ traits and parameterised by probabilities of transitions between these states. If our source data reveals two evolutionary transitions $111 \rightarrow 110$ and $110 \rightarrow 100$, (i) is a more likely parameterisation than (ii), as it supports evolutionary trajectories that are likely to give rise to those observations. HyperTraPS is used to calculate the associated likelihood, determining which parameterisation are accepted (perhaps (i)) and which are rejected (perhaps (ii)). (C) MCMC is used to build a posterior distribution of parameterisations based on the associated likelihood of observed transitions, producing an ensemble of possible evolutionary landscapes. (D) This posterior distribution is then summarised by recording the probability with which a given gene is lost at a given ordering on an evolutionary pathway. (E) Illustration of the distinct mitochondrial gene sets present in the source dataset, ordered vertically from highest to lowest gene content. Rows are genomes, columns are mitochondrial genes (an example species is given in grey for each genotype). Black and white pixels represent present and absent genes respectively. (F) A taxonomic relationship between organisms used in this study. Each leaf is an organism in which the set of present mitochondrial genes has been characterised. Coloured pairs of nodes denote those ancestor/descendant pairings where a change in mitochondrial gene complement is inferred to have occurred, with hue denoting the number of protein-coding genes present from blue (maximum 65) to red (minimum 3). Single letters in (E, F) denote positions of some well-known organisms by initial letter: *(H) Homo sapiens, (R) Reclinomonas americana, (P) Plasmodium vivax, (F) Fucus vesiculosus, (Z) Zea mays, (S) Saccharomyces cerevisiae, (A) Arabidopsis thaliana, (C) Caenorhabditis elegans, (D) Drosophila melanogaster.*

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for the parameterisation in Fig. 1 Bi, the probability of a given evolutionary process visiting state 111 and then state 110 is 0.98. But when L is higher and there are many different paths through which a given state may be visited (or avoided), it becomes prohibitively difficult to compute the probability of every appropriate path. Our approach provides an efficient way to compute the probability of observing a given pair of states, given a particular parameter set of intensities. Briefly, it accomplishes this by sampling the set of paths that start at the first state and finish at the second, but preferentially sampling the paths which are most likely. This preferential sampling is accomplished firstly by ensuring that no sampled paths involve transitions to states from which the second desired state cannot be reached, and secondly by choosing each step on a sampled path proportionally to its intensity. The amount of bias introduced by this preferential sampling is recorded at each step in the sampled path, and is corrected for to yield the probability required.

In more detail, we work with the 'evolutionary space' described in Methods, consisting of a set S of states comprising all possible patterns of presence and absence of L traits under consideration, and a set of transition probabilities $\pi_{s\to t}$ between each pair of states s and t, determining the stochastic dynamics of evolution (Fig. 1 A). To compute the probability of observing given evolutionary transitions, we require an estimate of the probability $P(s \to t|\pi)$ that an evolutionary trajectory on a network π passes through a 'target' state t, given that it passed through a 'source' point s. This probability is generally hard to compute for a given transition network due to the large number of possible paths that lead from s to t, since they may differ in any of L positions.

We show in the Supplementary Information that an estimate for $P(s \to t)$ can be efficiently computed. Consider a path c, consisting of a set of states beginning at s and ending at t, constructed as follows. In the ith path step c^i, identify the set $T(c^i)$ of all accessible states that are compatible with t (meaning that they do not lack genes that t possesses). Choose the next state c^{i+1} from $T(c^i)$ according to the associated transition probabilities normalised over $T(c^i)$ alone. The advantage of this sampling approach is that we can simulate trajectories guaranteed to start at s and end at t, thus avoiding wasting computational time on trajectories that do not contribute to the overall sum. We underline that our approach does not introduce bias between these pathways – it solely serves to focus computational energy on appropriate pathways. An estimate of $P(s \to t)$ is then efficiently given by averaging the quantity $\prod_i P(c^i \to \circ \in T(c^i))$ over a set of samples, where c^i is the ith state in path c, $\circ \in T(c^i)$ denotes any member of the set of states accessible from c^i that are compatible with final target t, and the sampling is taken over a set of paths, constructed according to the sampling scheme above, starting at s and ending at t (Fig. 6C). This estimate can be computed using the algorithm below, which to our knowledge has not been previously published; if this is true, we propose the name 'HyperTraPS', both standing for **hyper**cubic **tra**nsition **p**ath **s**ampling and referring to the act of forcing trajectories towards specific points on a hypercube. We note that HyperTraPS is in a sense more comparable to the Transition Interface Sampling or Forward Flux Sampling approaches found in statistical physics than Transition Path Sampling as understood in that context [36].

Algorithm 1. *Hypercubic transition path sampling (HyperTraPS)*

1. Initialise a set of N_h trajectories at s.
2. For each trajectory i in the set of N_h:
   (a) Compute the probability of making a move to a t-compatible next step (for the first step, all trajectories are at the same point and the probability for each is thus the same); record this probability as α_i'.
   (b) If current state is s, set $\alpha_i = \alpha_i'$, otherwise set $\alpha_i \to \alpha_i \alpha_i'$.
   (c) Select one of the available t-compatible steps according to their relative weight. Update trajectory i by making this move.
3. If current state is everywhere t go to 5, otherwise go to 2.
4. $\hat{P}(s \to t) = N_h^{-1} \sum_i \alpha_i$.

N_h, the number of sampled trajectories, is a parameter of the algorithm. Lower numbers will be computationally cheaper but will give a poorer sampling of possible trajectories and thus a less accurate estimate $\hat{P}(s \to t)$. In the Supplementary Information we give a more detailed description of HyperTraPS and show a range of validation checks for its functionality (Fig. 7A-D , 7E). To infer patterns of evolutionary dynamics with HyperTraPS, first a phylogeny is used to identify the evolutionary transitions that have occurred throughout history (Fig. 1 A and Supplementary Information). Then HyperTraPS is used to compute the joint probability of observing these transitions given a trial parameterisation π (Fig. 1 B). An MCMC approach is used to sample parameterisations which are associated with high likelihoods (Fig. 1 C), forming a posterior distribution over π which can be summarised and visualised (Fig. 1 D).

After addressing the specific question of mtDNA evolution, in the final results section we compare HyperTraPS qualitatively and quantitatively with existing approaches addressing the broad question of characterising the evolution of traits over a phylogeny.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

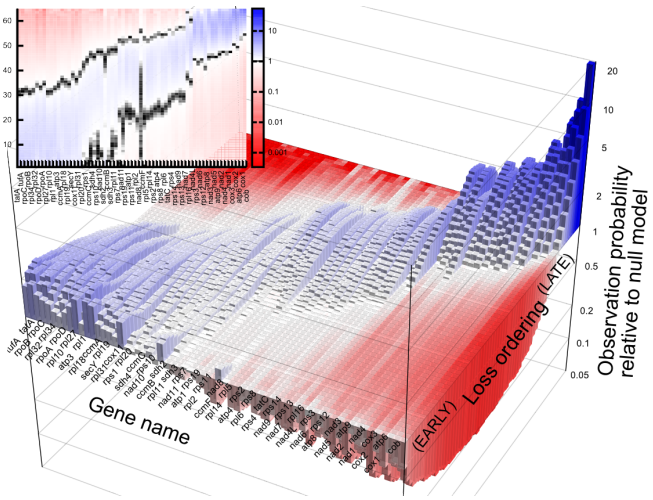

Figure 2: **The inferred ordering of mitochondrial gene loss is highly structured and non-uniform.** The probability that a given gene is lost at a given time ordering in the process of mitochondrial gene loss. The flat surface in the main plot and black contour in the inset give the probability $(1/L)$ associated with a null model where all genes are equally likely to be lost at all times. Blue corresponds to a probability above that expected from this null model; red corresponds to a probability below this null model. Genes are ordered by mean inferred loss time.

The inferred pattern of mitochondrial gene loss through evolution

We analyzed the complete annotated mitochondrial genomes of 2015 species from GOBASE [37], identifying the presence and absence of each *R. americana* mitochondrial gene for each genome. Across the 2015 species, 74 distinct combinations of genes were present (Fig. 1 D). Visible vertical clusters of retained genes include subunits of Complexes I (*nad[X]*), III (*cox[X]*) and V (*atp[X]*), and the small subunit of the mitochondrial ribosome (*rps[X]*). We then mapped the differences between each genome onto a phylogeny of all the species in our dataset (Fig. 1 E). We employ two assumptions about mtDNA evolution: first, that mitochondrial gene loss is rare, and second, that mitochondrial gene gain is negligible. These assumptions allow us to reconstruct the mitochondrial genomes at ancestral nodes in the phylogeny (Supplementary Information; Fig. 6A-B). To ensure that our approach was robust to potential errors in annotation and phylogeny construction, we repeatedly perturbed our source data and confirmed that our results were comparable across perturbed datasets (Fig. 7E ; Methods and Supplementary Information).

We used the HyperTraPS algorithm within a Bayesian MCMC framework to compute the probabilities of different patterns of mtDNA gene loss, given observed changes in mtDNA across the Tree of Life and uninformative uniform priors on transition probabilities. Fig. 2 shows a summary of these results, illustrating the probability with which a given gene is lost at a given step in time from a 'full' genome containing all genes found in *R. americana*, to an 'empty' mitochondrial genome containing no genes. The figure heuristically represents possible pathways for the evolutionary history of a 'typical' mitochondrial genome in an evolving eukaryotic lineage. The pattern of mitochondrial gene loss is remarkably structured and non-random, rejecting the possibility that the genes retained are shared across many species by chance. The inferred structure also quantitatively supports intuitive observations, for example, cytochrome b (*cob*) is observed in almost every known mitochondrial genome, whereas the secY-independent protein translocase component *tatA* is only observed in *R. americana*. We broadly observe three classes of genes: early loss (including Complex II *sdh[X]* genes, many ribosomal *rps[X]* and *rpl[X]* genes); intermediates present in a variety of taxa (including plants and fungi) but lost in animals (including more subunits of the mitochondrial ribosome and some Complex V *atp[X]* genes); and highly retained genes (including Complex I *nad[X]* genes, Complex IV *cox1-3* genes and cytochrome b *cob*). Different lineages have clearly experienced different specific gene loss trajectories (for example, *Schizosaccharomyces pombe* possesses *cox2* but not *cox3*, and *Babesia bovis* possesses *cox3* but not *cox2*, Fig. 1 B), but the probabilistic trend observed in Fig. 2 holds broadly across eukaryotes.

Gene loss in electron transport chain complex proteins

Our inferred order of gene loss shows that genes encoding distinct mitochondrial protein complexes were lost in different 'phases' throughout evolutionary time. As a first step in more detailed characterisation of this loss patterning, we sought to examine the order of gene loss within mitochondrial protein complexes to address a long-standing hypothesis for why some genes are retained in organelles. As described in the introduction, the CoRR hypothesis suggests that components of central importance to organelle function are preferentially encoded in the organelle genome, to allow localized control of the assembly of protein complexes [23]. This allows individual mitochondria to adjust the stoichiometry of ETC complexes in response to demands or stresses, without affecting the regulation of other mitochondria within the same cell. We reasoned that protein subunits occupying energetically central positions within their complexes likely exert the most control over complex assembly [38] and thus provide a means to test this hypothesis.

We analyzed the crystal structures that have been determined for ETC complexes II, III and IV (see Methods). We found that the genes that encode the proteins with the strongest binding energy (computed using PDBePisa; see Methods) within each complex have invariably been retained in mitochondrial genomes more commonly than those encoding other subunits of the complex (Fig. 3). This relationship is apparent across the entire eukaryotic Tree of Life. The two genes most

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

retained by mitochondrial genomes throughout life, *cob* and *cox1*, are the most energetically central components of their complexes (Complex III and Complex IV respectively). The *sdh[2-4]* genes controlling Complex II, and *cox[2-3]*, which play an important role in Complex IV but are less central than *cox1*, represent intermediate points on this spectrum. Many organisms do not encode any Complex II subunits in mtDNA: those that do so retain the energetically most central *sdh[2-4]* genes and not *sdh1*. *cox[2-3]* are retained in the mitochondria of most but not all organisms. By contrast, the energetically more peripheral gene products are invariably encoded by the nucleus. This link between biophysical and evolutionary features of mitochondrial genes is consistent with the CoRR hypothesis, supporting its applicability to organellar genomes of multiple and diverse taxa across the tree of life.

During the preparation of this report, crystal structure data for ETC complex I became available on the PDB [39], affording a valuable opportunity to independently test the prediction of our model. We computed interaction energies for Complex I subunits and found that protein products encoded by mtDNA genes also display the strongest interaction energies, consistent with the pattern we observed with other complexes (Fig. 3). As is the case with the other ETC complexes, the Complex I subunit with the strongest interaction energy, *nad1*, is one of the most commonly retained genes in mitochondrial genomes across all taxa.

GC content and protein hydrophobicity are both required to predict mitochondrial gene retention

We next sought to examine additional factors that predict the probability that a given gene is retained in the mitochondrial genome. To do this, we gathered data for ten properties of mitochondrial genes (or their encoded proteins) that have been hypothesized to influence the probability that a given gene is retained in the mitochondrial genome (Supplementary Table I). These hypotheses can be viewed as predicting a strong link between the corresponding gene property and the propensity with which a gene is lost from mtDNA. Our characterisation of the probability with which a gene is lost in a given ordering allows us to quantify the strength of these hypothesized links within a probabilistic framework.

We used a linear model framework to explore the ability of features of mitochondrial genes, or the proteins they encode, to predict the inferred patterns of gene loss shown in Fig. 2 . As described in Methods, we performed our Bayesian model selection approach with two datasets, one corresponding to genetic features in *R. americana* alone and one with features averaged across taxa. This approach allows us to control both for specific properties of *R. americana* and for cross-species variation. As shown in this section and the Supplementary Information, results from both datasets are very comparable, illustrating the robustness of our findings.

The resultant posterior probabilities over model structure displayed a striking favouring of GC content and protein product hydrophobicity as predictors of whether or not genes are retained in mitochondrial genomes; GC content (B) and hydrophobicity (C) appear in 97-98% of inferred model structures using both *R. americana* gene properties (Fig. 4 A) and averaged gene properties (Supplementary Information; Fig. 8A-C). Few other properties are so represented, with the exception of strandedness (J), which appears in 56% of inferred models using averaged gene properties (but few models using *R. americana* properties). Results involving uniform priors on model structure, and using an expanded set of features, were compatible with these findings (Fig. 8A-C ; Supplementary Information). The absence of other features in favoured models illustrates a lack of support for other retention determinants including, for example, features related to genetic code variability and mutational robustness. We did not find strong evidence that gene expression is directly related to mtDNA gene retention (Supplementary Information; Fig. 8D), although this analysis was based on a smaller dataset limited by available expression data (see Methods and Supplementary Information).

Our analysis provides strong support for genes with high GC content *relative to other genes in the same organism* being preferentially retained in mtDNA. As we discuss in detail in the Supplementary Information, patterns of GC content vary dramatically between species (Fig. 9 ; [27]), but this *inter*-species variability is never incompatible with the *intra*-species favouring of GC-rich genes. Notably, at a sequence level, we have found that protein-coding genes in mtDNA can in some species display a bias *against* GC content, in the sense that GC-poor codons are used more often than GC-rich codons to encode a given amino acid (Supplementary Information; Fig. 9). This is most likely due to the strong asymmetric mutational pressure arising from the hydrolytic deamination of cytosine into uracil, which is converted into thymine. This mutational pressure likely enriches GC-poor codons in organellar genomes [26]. Our results suggest a tension between this 'entropic' mutational drive at the sequence level (decreasing genome-wide GC content) and a selective drive at the genomic level (retaining genes with higher GC content).

To further elucidate the nature of this tension, we examined the GC content of individual genes at nonsynonymous and synonymous positions in *R. americana* and averaged across taxa. We found that GC content was lower at synonymous positions than at nonsynonymous positions in *R. americana*, but that this difference largely disappeared in the taxa-averaged data (Fig. 9B). This difference in *R. americana* is consonant with a link between GC content and structural conservation: nonsynonymous sites retain their GC content as conser-

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

vation pressure balances the asymmetric mutation pressure, while synonymous sites are free to decrease GC content in response to asymmetric mutation. In this picture, the link between GC content and gene retention could arise indirectly through another relationship: genes most retained in mtDNA being most highly conserved (perhaps due to their structural importance in protein complex assembly, as above). However, the absence of this signal in the taxa-averaged data suggests that this relationship may not hold across all the species we consider. We also tested the link with conservation by using nonsynonymous, synonymous, and total GC content as different terms in our model selection process (Supplementary Information, Fig. 9C). The model selection process always favoured total GC content, suggesting that the conservation link, while explanatory in *R. americana*, does not represent the only way in which GC content is linked to gene retention (see Discussion).

The parameterisation of the most common model always favoured the retention of genes encoding proteins with high hydrophobicity and high GC content. In Fig. 4 B we illustrate a specific parameterisation of this model, chosen to optimise the least-squares difference between model prediction and mean inferred ordering (from Fig. 2). This parameterisation shows a strong gene-by-gene correlation with the mean inferred loss ordering, with Pearson's $R^2 = 0.51$ ($p \simeq 2 \times 10^{-11}$, see Supplementary Information for interpretation) and Spearman's $\rho = 0.72$. The model also strongly correlated with observation frequencies of mtDNA genes in the original dataset (Fig. 1 B, 8G).

Intuitively, it may be expected that the signals associated with GC content, hydrophobicity, and energetic centrality may reflect different aspects of the same fundamental feature: GC-rich codons encode hydrophobic amino acids, and hydrophobic proteins are likely to occupy 'core' positions in complexes and within membranes. However, there is evidence that these features are at least somewhat decoupled. Fig. 8F shows an absence of strong correlations between the three features for the genes we consider: Complexes II and IV display a moderate correlation between scaled energetic centrality and GC content (though sample size is limited by the small number of subunits), but no relationship exists for other complexes or hydrophobicity. Additionally, our model selection process will discard redundant information; if all the predictive power associated with GC content was already present in the hydrophobicity data, GC content would not be identified as a joint factor in the selected models. While the links above are certainly true in some contexts, we cannot avoid the conclusion that independent features associated with GC content, hydrophobicity, and energetic centrality all contribute to gene loss propensity. A combination of these, as opposed to any single feature, is therefore required to account for observed patterns of mtDNA gene retention.

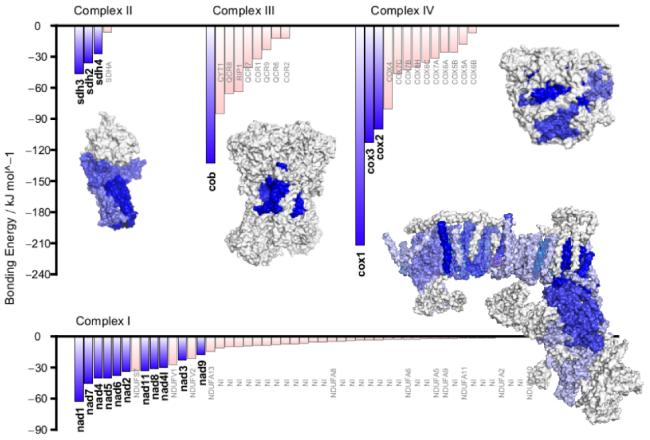

Figure 3: **Assembly centrality, gene retention, and co-localisation of redox regulation.** Interaction energy of subunits within Complexes I-IV computed with PDBePISA (see Methods). Blue bars denote subunits encoded by mtDNA in at least one eukaryotic species; pink bars denote subunits always encoded in the nucleus. Inset crystal structures show the location of mtDNA-encoded subunits in blue (darker shades show higher interaction energies).

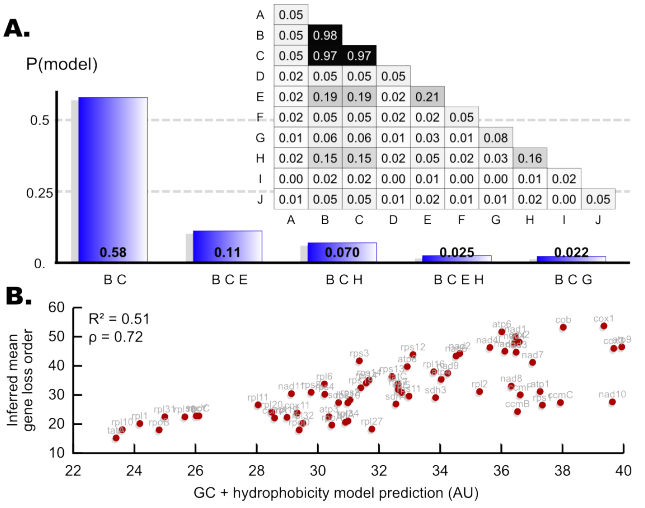

Figure 4: **Features predicting mitochondrial gene retention.** (A) Bars show posterior probabilities for individual models for mitochondrial gene loss based on gene properties in *R. americana*, given the inferred gene loss patterns and a prior favouring parsimonious models. Inset matrix show the posterior probability with which features are present in these models for gene loss (diagonal elements correspond to a single feature, off-diagonal elements to a pair of features). Labels A-J denote model features as described in Methods; B is GC content and C is protein product hydrophobicity. (B) Predicted loss ordering from GC & hydrophobicity model (horizontal axis) against inferred mean loss ordering (vertical axis).

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

HyperTraPS: Comparison with related approaches, and future applicability

The study of the evolution of traits across phylogenetically related lineages has an extensive history and associated literature (reviewed in Ref. [40]). Methods have been developed in this field to infer phylogenetic trees, to infer the evolutionary dynamics of traits on phylogenies, and to jointly infer both (a variety of each type are included in Ref. [41]). Often a single (continuous or discrete) trait or pairs of traits are considered; one approach that is particularly notable in the context of this study employs reversible-jump MCMC to explore the potentially correlated evolution of two discrete traits on a phylogeny [42]. However, we are unaware of an approach in this field that can consider the evolution of a large ($L = 65$) set of potentially interacting traits on a phylogeny. These approaches provide alternative and sophisticated approaches for characterising uncertainty in phylogenies. However, we employ our straightforward and robust approach (see Methods) due to the taxonomic range of our study (and corresponding difficulty in using e.g. particular features like sequence alignments for phylogenies), our interest in an ordering of trait changes rather than an absolute temporal dimension, and our focus on trait dynamics rather than phylogenetic structure *per se*, which mtDNA studies have previously addressed [32].

A rather disconnected branch of literature attempts to describe the (usually irreversible) acquisition of coupled binary traits with time, usually given a set of independent observations of this process. This field is largely motivated by the target of inferring the dynamics of cancer progression, with the binary traits under consideration often corresponding to the presence or absence of chromosomal aberrations. Refs. [43] and [44] review and classify approaches to this question, which often involve assumptions about the causal links between traits (for example, that the network of trait acquisition influence is tree-like). Our method allows these assumptions to be relaxed to the case where traits can influence acquisitions arbitrarily, and in concert allows posterior ordering to be derived, facilitating subsequent exploration of connected factors.

The method proposed in Ref. [45] employs a similar Markov chain philosophy (and indeed uses the same strategy for reducing parameter space as in our method). This approach can be adapted to mirror the same likelihood function as HyperTraPS, by consider each transition between our mtDNA states as a start and end observation of traits. However, the likelihood calculation proposed in that study has a complexity of $O(nL2^k)$, where n is the number of observations, L the number of traits, and k a number of trait differences that separate two observations. As, in this study, k is sometimes of the same order of L, and $L = 65$, this likelihood calculation is intractable. Another approach, 'phenotypic landscape inference' [46], uses likelihoods estimated in time $O(nL^2r)$ (where r is the number of paths used to characterise a given transition), using an unbiased search which necessitated $r \sim O(2^L)$ for satisfactory convergence (and many likelihood calculations to satisfactorily explore the parameter space of size 2^L), and so is also intractable for $L = 65$. By comparison, the efficient probability-weighted algorithm in HyperTraPS has a complexity of $O(nk^2r)$, and the resultant polynomial rather than exponential scaling with k means that HyperTraPS provides a tractable (approximate) likelihood where the approaches of Refs. [45] and [46] would be intractable, making the investigation of these larger questions computationally feasible for the first time. Polynomial rather than exponential scaling in the number of traits means that progression dynamics can now be inferred for problems involving many traits without necessitating assumptions of tree or other structures underlying trait relationships (although such assumptions can straightforwardly be included by applying restrictive priors to the appropriate transition intensities). Additionally, as we demonstrate, increased computational speed facilitates fully Bayesian analyses of the dynamics of evolutionary/progression systems, allowing subsequent analysis of explanatory features and mechanisms, probabilistic statements about expected progression pathways from a given state, predictions about unmeasured trait values, and other important deliverables.

Fig. 7F-H illustrates several existing approaches applied to our data on mtDNA gene loss. First, the 'oncotrees' package, which attempts to infer a tree of relations between different traits that are acquired with time [47]. The trees that emerge from this approach are heuristically comparable to our findings: *cob* is notable at one limit of loss dynamics, with *cox[X]*, *nad[X]*, and some *atp[X]* genes occupying points late in the ordering hierarchy, whereas rare genes like *secY* and *tatA* are present far lower. This picture provides independent confirmation for our findings, but does not characterise explicit posterior probabilities in the detail that our HyperTraPS-led approach does, and would not admit further analysis within the same framework to determine features that predict loss propensity. Second, the 'OrderMutation' approach of Ref. [48], which attempts to infer the probability with which a given trait changes at a given ordering. This approach produces results comparable to ours when applied to an $n = 10$ subset of the genes we consider, but struggles with a larger $n = 20$ set and fails to process larger datasets. Third, the 'simmap' approach, which characterises the transitions underlying stochastic transitions between different trait patterns on a given phylogeny. This approach (when restricted to consider gene loss as irreversible) generates plausible realisations of evolutionary trajectories for small subsets of genes on our phylogeny (as illustrated in Fig. 7F-H) and can characterise uncertainty in this trajectories. However, the 2^L different potential trait states in our

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

mtDNA system, most of which are not observed in contemporary samples, rendered this approach intractable.

We believe that our highly generalizable mathematical approach may be used to answer a broad array of questions in biology and medicine, offering the ability to work with a general search space comprising dozens of features, which are unrestricted and may comprise biologically distinct classes (for example, combining physical and genetic characteristics [46], or unrelated clinical phenotypes), and the ability to make probabilistic predictions with quantified uncertainty about behaviour given a particular state, and unmeasured features of samples (verified in a plant study in Ref. [46]). Our approach can be employed in cases where discrete traits change irreversibly with time (though we are pursuing the extension to reversible transitions), and we anticipate it being of use in fields (with corresponding traits in parentheses) including cancer progression [44] (chromosomal aberrations); antibiotic resistance in tuberculosis [49] (resistance to given drugs); chloroplast evolution [50] ('tribes' of organelle genes); and paleontology (discrete morphological traits).

Discussion

We have developed new mathematical machinery combining stochastic modelling with Bayesian inference to infer structure and variability of evolutionary pathways. Our approach is potentially a widely-applicable tool for inferring the histories and predicting likely trajectories of any evolutionary system involving the loss or acquisition of distinct characteristics. As we discuss above, we anticipate that the power of HyperTraPS to address and quantify uncertainty in evolutionary and progression dynamics will be of use in a considerable range of future research questions in biology and medicine. Applied to the question of mtDNA gene loss, we have identified striking structure in mitochondrial evolution across eukaryotes, involving parallel losses across lineages and fine-grained lineage variation imposed on a broad, predictable cross-species trend.

We have focussed on protein-coding mitochondrial genes, to facilitate an analysis of the potential modulating factors of this gene loss. The mitochondrial genes encoding, for example, rRNA and tRNA are likely subject to different evolutionary pressures and will be the target of future investigation. Considering mtDNA genes coding for electron transport chain proteins, we have quantitatively demonstrated that genes central to the assembly of ETC protein complexes are preferentially retained in eukaryotic mtDNA. This observation suggests a picture of a controllable subset of 'core' genes retained by the mitochondrion to allow localised bioenergetic modulation, with a cytoplasmic pool of 'periphery' subunits ready to assemble around any newly-produced core, congruent with recent findings that genes encoding central subunits of the mitochondrial ribosome are preferentially retained in mtDNA [24]. This idea of a localised and controlling relationship between mtDNA and energetic machinery supports the central aspects of the CoRR hypothesis [23] and the picture of locally addressable 'quality control' of mitochondria [51, 52]. A related recent hypothesis [6] proposes that the acquisition and subsequent genetic reduction of mitochondria was a major facilitating step in the evolution of complex life, as localised cellular power stations controlled through a small number of genes dramatically increase a cell's available energy per gene, allowing genetic exploration without sacrificing energy. Our identification of energetically central 'control' genes as preferentially retained in mitochondrial genomes is consistent with CoRR across the range of taxa that we consider.

Energetic centrality in ETC complexes provides a suggestion towards why a particular subset of the thousands of genes possessed by 'proto-mitochondria' has been retained in modern eukaryotes. A related question is why, within this subset, certain genes are retained in the mtDNA of more species than others. Considering the full range of mtDNA-encoded genes, we have identified two features – GC content, and the hydrophobicity of an encoded protein product – that strongly predict the propensity of a gene to be retained in mtDNA across taxa. Other proposed features modulating retention, such as genetic code differences, gene expression levels, and GC skew have much lower support from our approach, and a link between mtDNA retention and gene expression levels is evident in some but not all species, and likely secondary to the link with GC content (Supplementary Information). We underline that the explanatory contributions of GC content, hydrophobicity, and energetic centrality to gene retention are at least partly independent: each factor has individual explanatory power, suggesting mechanisms associated with each. As discussed by Allen and Cavalier-Smith following Ref. [53], a combination of features is therefore required to account satisfactorily for gene retention patterns, and our results quantitatively describe the explanatory components of this combination for the first time.

If a coupled picture where these features report the same underlying property – hydrophobic, central, GC-richly-encoded subunits obeying CoRR – is not the whole story, what other mechanisms could underlie the observed links? Regarding hydrophobicity, it has been suggested that import of hydrophobic proteins through membranes into the mitochondrion may be limiting [20, 14], though this picture is debated [23]. Additionally, it has been proposed that hydrophobic proteins are more likely to be directed to the endoplasmic reticulum than to the mitochondria when translated in the cytoplasm [21]. Importantly, a recent study has experimentally verified that mitochondrially-encoded proteins are indeed localized to the endoplasmic reticulum when expressed in HeLa cells [22] – thus providing a mechanism for hydrophobic modulation of mtDNA gene retention

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

without invoking the controversial membrane-traversal picture. Our results are consistent with the plausible picture that emerges – that the sub-cellular targeting of hydrophobic proteins is an important selective constraint that has favored retention of some genes in mitochondrial genomes.

The independent contribution of GC content to mtDNA gene retention is less mechanistically clear. We show in Figs. 9B and 9C that a putative link between structural conservation and gene retention may account for some, but not all, of the explanatory power of GC content. We suggest that the remaining connection between total GC content and mtDNA gene retention may be manifest through the physical and chemical properties of mtDNA and derived nucleic acids, and their survival in the damaging environment of the mitochondrion. Firstly, GC content has been hypothesised to modulate longevity through its influence on the thermodynamic stability of the mtDNA molecule [27]: nucleic acids containing more (stronger) GC bonds may be less prone to spontaneous 'bubble' formation and thus more protected from environmental mutagens. Secondly, adenine depletion has sometimes been observed during oxidative stress [54, 55]. MtDNA could thus preferentially utilise GC-rich codons to optimise chemical stability of nucleic acids in the mitochondrion, and to avoid depleting these redox-linked pools under stress, resulting in selection for increased GC content. This hypothesis is experimentally testable, for example by measuring the extent of DNA damage in GC-rich versus GC-poor mitochondrially encoded genes under stress, which we would expect to be exacerbated in mutants for mitochondrial DNA repair pathways.

Our findings give rise to further predictions on cellular and evolutionary scales. At a systematic level, observation of an organism with mtDNA encoding GC-poor, low-hydrophobicity, energetically peripheral genes and not GC-rich, hydrophobic, energetically central genes would stand against our theory. On a cellular level, our theory predicts that attempts to encode GC-rich, hydrophobic, energetically central genes in the nucleus rather than in mtDNA will adversely affect cell functionality. This prediction can be tested through artificial gene transfer experiments, and indeed, several recent and intriguing experimental studies in *S. cerevisiae* attempting to artificially transfer mtDNA genes to the nucleus allow us to test our theory. To our knowledge, such experiments have been attempted for genes *cob* [56] (unsuccessful), *atp9* [57] (unsuccessful but introduction of a nuclear-encoded version from another species succeeded), *cox2* [58] (successful after a small structural modification), *atp8* [59] (successful), and *rps3* (successful) [60]. This relative ordering of difficulty matches the predictions made from our theoretical treatment as observed in Fig. 5 , supporting our GC/hydrophobicity theory. Our individual mechanistic hypotheses can also be tested: for example, if redox-linked adenine depletion does constitute an important pressure favouring GC content, GC-richer mtDNA haplotypes should experience an advantage under oxidative stress. A confirmatory assay could be performed in several existing models where two different mtDNA haplotypes exist in admixture [61]. In addition to providing the first statistically robust support for long-debated hypotheses regarding mitochondrial gene loss, and identifying the combination of factors governing this process, our study thus also yields simple, experimentally testable predictions from a complex set of evolutionary transitions.

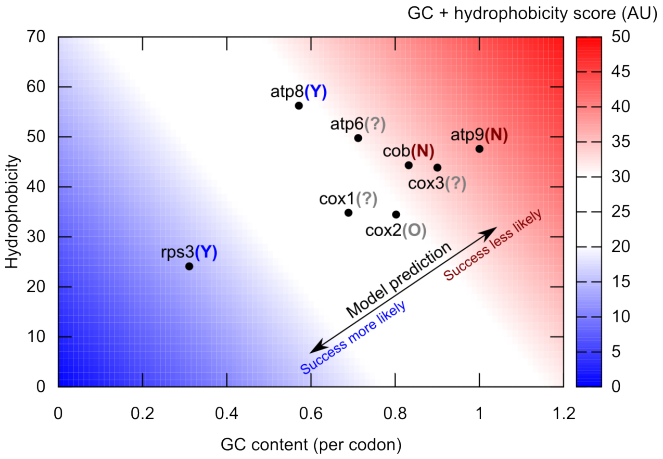

Figure 5: **Predicted and observed feasibility of experimental mito-nuclear gene transfer in *S. cerevisiae*.** The protein-coding mtDNA genes in *S. cerevisiae*, plotted on a space of GC content and protein hydrophobicity. Heat map gives the value of the fitted model from Fig. 4 B. Parentheses give the experimental status of mito-nuclear transfer for the given gene: Y – successful, N – unsuccessful; O – partial success (after structural modification), ? – currently unattempted. The success of transfers follows the model prediction for loss propensity, and predictions for the ease of unattempted gene transfers can be formulated.

Experimental Procedures

Data mining and curation for mtDNA gene content. The jakobid protozoan *Reclinomonas americana* has a large mitochondrial genome [9] which includes orthologues of every gene found in any organism's mitochondrial genome (with very few exceptions [62]). We use the set of $L = 65$ identified protein-coding genes in the mtDNA of *R. americana* as our reference set (Supplementary Information). We represent mitochondrial genomes from across all kingdoms of eukaryotic life as strings of L presence or absence markers, one for each protein-coding gene in the *R. americana* genome. To obtain these representations, we use the full set of data from GOBASE, the organellar genome database [37], consisting of annotated genome records for 2 015 species after duplicates were removed. We used the NCBI Taxonomy tool [63] to build an estimated tree of taxonomic relationships between all species recorded in the dataset, and manually verified the topology of this tree with comparison to the Encyclopedia of Life [64] and Tree of Life [65] projects. We then identify the changes that mitochondrial genome content has undergone throughout evolutionary history by comparing inferred ancestral mitochondrial properties with descendant properties (Fig. 6A-B ; Supplementary Information). To ensure that our results are robust with respect to perturbations in the details of the constructed phylogeny and this inference protocol, we applied sets of random changes to the resultant dataset and verified that our results were consistent across this set of random

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

changes (Supplementary Information).

Representing evolutionary space and transition networks. We consider an underlying 'evolutionary space' described by a transition network $\pi_{S_i \to S_j}$, giving the probability of a transition from state S_i to S_j, where states correspond to binary strings of length L as above. In any state, we consider transitions corresponding to the loss of exactly one mtDNA gene, restricting π to a hypercubic structure (Fig. 1 B). We write π using $\pi_{s \to t} = Z^{-1} P(\text{lose trait } i | \text{state } s) I(s,t,i)$, where $I(s,t,i)$ is an indicator function returning 1 if t is equivalent to s after the loss of trait i, and 0 otherwise, and Z is a normalising factor to ensure $\sum_t \pi_{s \to t} = 1$. This structure allows us to coarse-grain the parameterisation of the problem by writing $P(\text{lose trait } i | \text{state } s) = \exp\left(m_{i,1} + \sum_j m_{i,j+1}(1 - s_j)\right)$, where m is an $L \times (L+1)$ matrix m, with elements in the first row $m_{i,1}$ representing the 'default' probability of losing trait i, and elements in the subsequent rows $m_{i,j+1}$ describing how this default probability changes in a state where trait j has already been lost. This representation, as discussed in Ref. [45] where this philosophy is also employed, allows us to use $O(L^2)$ parameters rather than the full $O(2^L)$ set of transition probabilities, while retaining the ability to model both independent gene loss propensities and potential contingencies of the loss of one gene on the presence of others. We illustrate its ability to satisfactorily capture evolutionary behaviour in the Supplementary Information.

Inferring mtDNA gene loss ordering. We developed an extended and generalised version of phenotype landscape inference [46], using an algorithm we call HyperTraPS (hypercubic transition path sampling; see Results). We use HyperTraPS to compute the probability of observing the set of mitochondrial gene transitions that we infer from genomic data, given the matrix m (see above) representing transition probabilities between different mtDNA states. This probability is used in a Bayesian MCMC algorithm (Fig. 1 A-D; see Supplementary Information for quantitative details) to obtain a posterior distribution on m. This posterior distribution is summarised as a posterior distribution over mtDNA gene loss orderings (Fig. 2).

Bayesian model selection for features determining mtDNA gene loss propensity. To investigate the potential determining factors that govern mitochondrial gene loss ordering, we compiled a list of many physical and genetic features of mitochondrial genes using sequence data [37] and standard chemical data sources [66] (Supplementary Information). Features included (A) gene length, (B) GC content, (C) hydrophobicity of product, (D) molecular weight, (E) pK_a, (F) energetic cost of production [67], (G) codon universality (the proportion of codons whose interpretation varies across eukaryotes [14]), (H) mutational robustness [13], (I) GC skew, and (J) the strand on which the gene was encoded. As described in the Supplementary Information, this set of features allows us to explore existing hypotheses regarding mitochondrial gene retention (including, for example, hydrophobicity and genetic code differences), while also investigating gene properties that have not previously been explicitly considered in the literature (including GC content and an associated role for nucleic acid stability [27]). We also used RNASeq data from two species with $\sim$ 30 mtDNA genes (*Lolium perenne L.* [68] and *Phoenix dactylifera L.* [69]) to quantify links between gene retention and expression levels. These species have, to our knowledge, the highest number of mtDNA genes of any species in which gene expression has been quantified; we therefore selected them in order to quantify expression levels (and thus explore potential correlations) for the maximum possible number of mtDNA genes. Full details of how each feature is represented, and their links to individual hypotheses, is presented in the Supplementary Information. For A-J, we created two libraries of genetic properties for protein-coding mtDNA genes: one based upon values derived from *R. americana* (to control against interspecies variation of these properties) and one based upon values averaged over a set of species representing the full range of available eukaryotic diversity (to control against features specific to *R. americana*). We used a Bayesian model selection approach to compare the support of linear models involving combinations of these features given the patterns of mitochondrial gene loss inferred by HyperTraPS. We used two different classes of prior: first, a uniform prior probability over all possible models regardless of the number of features of each, and second, a prior probability exponentially decreasing with the number of features involved, to favor sparser and more parsimonious models.

Energetic centrality of ETC complex subunits. To explore the CoRR hypothesis, we investigated the interaction energy associated with a subset of mitochondrial genes in their respective protein complexes, as a measure of the energetic centrality that these genes play in the assembly and structure of each complex [38]. To quantify this measure, we used the PDBePISA tool [70, 71] to analyse the energetic interactions of the subunits in solved structures of electron transport chain complexes in the PDB [72] (specifically, Complex II (PDB 2h88, *Gallus gallus*), Complex III (PDB 3cxh, *Saccharomyces cerevisiae*, Complex IV (PDB 1oco, *Bos taurus*), and recently Complex I (PDB 4uq8, *Bos taurus*)), and identified the subunits corresponding to genes that are present in our reference set, using annotations associated with each PDB entry, and the UniProt resource [73].

Acknowledgments

The authors thank A. Earl, N. Jones, A. Monti, and E. Røyrvik for valuable discussions.

References

[1] L. Margulis. *Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth.* Yale University Press, 1970.

[2] M. Gray. Mitochondrial evolution. *Cold Spring Harbor Perspectives In Biology*, 4:a011403, 2012.

[3] M. Odintsova and N. Yurina. Genomics and evolution of cellular organelles. *Russian Journal Of Genetics*, 41:957, 2005.

[4] T. Kleine, U. Maier, and D. Leister. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual Review Of Plant Biology*, 60:115, 2009.

[5] G. Burger, M. Gray, and B. Franz Lang. Mitochondrial genomes: anything goes. *Trends In Genetics*, 19:709, 2003.

[6] N. Lane and W. Martin. The energetics of genome complexity. *Nature*, 467:929, 2010.

[7] B. Boussau, E. Karlberg, A. Frank, B. Legault, and S. Andersson. Computational inference of scenarios for α-proteobacterial genome evolution. *Proceedings Of The National Academy Of Sciences*, 101:9722, 2004.

[8] A. Vaidya and M. Mather. Mitochondrial evolution and functions in malaria parasites. *Annual Review Of Microbiology*, 63:249, 2009.

[9] B. Lang et al. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature*, 387:493, 1997.

[10] D. Wallace and D. Chalkia. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor Perspectives In Biology*, 5:a021220, 2013.

[11] S. Dean, M. Gould, C. Dewar, and A. Schnaufer. Single point mutations in atp synthase compensate for mitochondrial genome loss in trypanosomes. *Proceedings of the National Academy of Sciences*, 110:14741, 2013.

[12] C. Saccone et al. Evolution of the mitochondrial genetic system: an overview. *Gene*, 261:153, 2000.

[13] J. Blanchard and M. Lynch. Organellar genes: why do they end up in the nucleus? *Trends In Genetics*, 16:315, 2000.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[14] K. Adams and J. Palmer. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics And Evolution*, 29:380, 2003.

[15] T. Cavalier-Smith. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biology Letters*, 6:342, 2010.

[16] E. Hazkani-Covo, R. Zeller, and W. Martin. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics*, 6:e1000834, 2010.

[17] W. Ford Doolittle. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends In Genetics*, 14:307, 1998.

[18] J. Allen and J. Raven. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. *Journal Of Molecular Evolution*, 42:482, 1996.

[19] D. Daley and J. Whelan. Why genes persist in organelle genomes. *Genome Biology*, 6:1, 2005.

[20] J. Popot and C. Vitry. On the microassembly of integral membrane proteins. *Annual Review Of Biophysics And Biophysical Chemistry*, 19:369, 1990.

[21] G. von Heijne. Why mitochondria need a genome. *FEBS letters*, 198(1):1–4, 1986.

[22] P. Björkholm, A. Harish, E. Hagström, A. Ernst, and S. Andersson. Mitochondrial genomes are retained by selective constraints on protein targeting. *Proceedings of the National Academy of Sciences*, 112(33):10154–10161, 2015.

[23] J. Allen. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proceedings of the National Academy of Sciences*, page 201500012, 2015.

[24] U. Maier et al. Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. *Genome Biology And Evolution*, 5:2318, 2013.

[25] W. Martin and C. Schnarrenberger. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Current Genetics*, 32:1, 1997.

[26] A. Reyes, C. Gissi, G. Pesole, and C. Saccone. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology And Evolution*, 15:957, 1998.

[27] D. Samuels. Life span is related to the free energy of mitochondrial DNA. *Mechanisms Of Ageing And Development*, 126:1123, 2005.

[28] B. Nabholz, H. Ellegren, and J. Wolf. High Levels of Gene Expression Explain the Strong Evolutionary Constraint of Mitochondrial Protein-Coding Genes. *Mol. Biol. Evol.*, 30:272, 2012.

[29] E. Desmond, C. Brochier-Armanet, P. Forterre, and S. Gribaldo. On the last common ancestor and early evolution of eukaryotes: reconstructing the history of mitochondrial ribosomes. *Research In Microbiology*, 162:53, 2011.

[30] M. Bernt et al. Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. *Molecular Phylogenetics And Evolution*, 69:320, 2013.

[31] J. De Las Rivas, J. Lozano, and A. Ortiz. Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Research*, 12:567, 2002.

[32] S. Kannan, I. Rogozin, and E. Koonin. Mitocogs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. *BMC evolutionary biology*, 14(1):237, 2014.

[33] B. Scheel and B. Hausdorf. Dynamic evolution of mitochondrial ribosomal proteins in holozoa. *Molecular phylogenetics and evolution*, 76:67, 2014.

[34] P. Keeling et al. The tree of eukaryotes. *Trends In Ecology & Evolution*, 20:670, 2005.

[35] K. Adams, Y. Qiu, M. Stoutemyer, and J. Palmer. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proceedings Of The National Academy Of Sciences*, 99:9905, 2002.

[36] R. Allen, C. Valeriani, and P. ten Wolde. Forward flux sampling for rare event simulations. *Journal of Physics: Condensed Matter*, 21(46):463102, 2009.

[37] E. OBrien et al. GOBASE: an organelle genome database. *Nucleic Acids Research*, 37:D946, 2009.

[38] E. Levy, E. Erba, C. Robinson, and S. Teichmann. Assembly reflects evolution of protein complexes. *Nature*, 453:1262, 2008.

[39] K. Vinothkumar, J. Zhu, and J. Hirst. Architecture of mammalian respiratory complex I. *Nature*, 515:80, 2014.

[40] B. O'Meara. Evolutionary inferences from phylogenies: a review of methods. *Annual Review of Ecology, Evolution, and Systematics*, 43:267–285, 2012.

[41] L. Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[42] M. Pagel and A. Meade. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. *The American Naturalist*, 167(6):808–825, 2006.

[43] N. Beerenwinkel, R. Schwarz, M. Gerstung, and F. Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[44] L. Loohuis et al. Inferring tree causal models of cancer progression with probability raising. *PLoS ONE*, 9:e108358, 2014.

[45] M. Hjelm, M. Höglund, and J. Lagergren. New probabilistic network models and algorithms for oncogenesis. *Journal of Computational Biology*, 13(4):853–865, 2006.

[46] B. Williams, I. Johnston, S. Covshoff, and J. Hibberd. Phenotypic landscape inference reveals multiple evolutionary paths to C4 photosynthesis. *Elife*, 2, 2013.

[47] A. Szabo and L. Pappas. Oncotree: Estimating oncogenetic trees. 2013.

[48] A. Youn and R. Simon. Estimating the order of mutations during tumorigenesis from tumor genome sequencing data. *Bioinformatics*, 28(12):1555–1561, 2012.

[49] A. Izu, T. Cohen, and V. DeGruttola. Bayesian estimation of mixture models with prespecified elements to compare drug resistance in treatment-naive and experienced tuberculosis cases. *PLoS Computational Biology*, 9(3):e1002973, 2013.

[50] L. Cui, N. Veeraraghavan, A. Richter, K. Wall, R. Jansen, J. Leebens-Mack, I. Makalowska, et al. Chloroplastdb: the chloroplast genome database. *Nucleic acids research*, 34:D692, 2006.

[51] T. Tatsuta and T. Langer. Quality control of mitochondria: protection against neurodegeneration and ageing. *The EMBO Journal*, 27:306, 2008.

[52] G. Twig, B. Hyde, and O. Shirihai. Mitochondrial fusion, fission and autophagy as a quality control axis: the bioenergetic view. *Biochimica Et Biophysica Acta (BBA)-Bioenergetics*, 1777:1092, 2008.

[53] J. Allen. The function of genomes in bioenergetic organelles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1429):19–38, 2003.

[54] T. Aalto and K. Raivio. Nucleotide depletion due to reactive oxygen metabolites in endothelial cells: effects of antioxidants and 3-aminobenzamide. *Pediatric Research*, 34(5):572–576, 1993.

[55] M. Ott, V. Gogvadze, S. Orrenius, and B. Zhivotovsky. Mitochondria, oxidative stress and cell death. *Apoptosis*, 12(5):913–922, 2007.

[56] M. Claros et al. Limitations to in vivo import of hydrophobic proteins into yeast mitochondria. *European Journal Of Biochemistry*, 228:762, 1995.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[57] M. Bietenhader et al. Experimental relocation of the mitochondrial ATP9 gene to the nucleus reveals forces underlying mitochondrial genome evolution. *PLoS Genetics*, 8:e1002876, 2012.

[58] L. Supekova, F. Supek, J. Greer, and P. Schultz. A single mutation in the first transmembrane domain of yeast COX2 enables its allotopic expression. *Proceedings Of The National Academy Of Sciences*, 107:5047, 2010.

[59] P. Nagley et al. Assembly of functional proton-translocating ATPase complex in yeast mitochondria with cytoplasmically synthesized subunit 8, a polypeptide normally encoded within the organelle. *Proceedings Of The National Academy Of Sciences*, 85:2091, 1988.

[60] M. Sanchirico et al. Relocation of the unusual VAR 1 gene from the mitochondrion to the nucleus. *Biochemistry And Cell Biology*, 73:987, 1995.

[61] J. Burgstaller, I. Johnston, and J. Poulton. Mitochondrial dna disease and developmental implications for reproductive strategies. *Molecular human reproduction*, 21(1):11–22, 2015.

[62] G. Pont-Kingdon et al. Mitochondrial DNA of the coral Sarcophyton glaucum contains a gene for a homologue of bacterial MutS: a possible case of gene transfer from the nucleus to the mitochondrion. *Journal Of Molecular Evolution*, 46:419, 1998.

[63] E. Sayers et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39:D38, 2011.

[64] C. Parr et al. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal*, 2, 2014.

[65] The Tree of Life Web Project. The Tree of Life Web Project, 2007.

[66] D. Lide. *Handbook of chemistry and physics*. CRC Press, 1991.

[67] M. Barton, D. Delneri, S. Oliver, M. Rattray, and C. Bergman. Evolutionary systems biology of amino acid biosynthetic cost in yeast. *PloS One*, 5:e11935, 2010.

[68] M. Islam et al. The genome and transcriptome of perennial ryegrass mitochondria. *BMC Genomics*, 14:202, 2013.

[69] Y. Fang et al. A complete sequence and transcriptomic analyses of date palm (phoenix dactylifera l.) mitochondrial genome. *PLoS ONE*, 7:e37164, 2012.

[70] E. Krissinel and K. Henrick. Inference of macromolecular assemblies from crystalline state. *Journal Of Molecular Biology*, 372:774, 2007.

[71] S. Velankar et al. PDBe: protein data bank in Europe. *Nucleic Acids Research*, page gkp916, 2009.

[72] H. Berman et al. The protein data bank. *Nucleic Acids Research*, 28:235, 2000.

[73] UniProt Consortium et al. The universal protein resource (UniProt). *Nucleic Acids Research*, 36:D190, 2008.

[74] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings Of The IEEE*, 77:257, 1989.

[75] J. Bollback. Simmap: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7:88, 2006.

[76] O. Monera, T. Sereda, N. Zhou, C. Kay, and R. Hodges. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *Journal Of Peptide Science*, 1:319, 1995.

[77] C. Craig and R. Weber. Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by Escherichia coli. *Molecular Biology And Evolution*, 15:774, 1998.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Information

Mitochondrial gene set acquisition and curation

The GOBASE database was use to obtain a set of records of all recorded mitochondrial genomes. Each record contained a species name, accession ID, and set of mitochondrial genes present. In the vast majority of cases, these sets formed a subset of the genes present in *Reclinomonas americana*, our reference organism. After duplicates were removed, our dataset included the genomes of $2,015$ distinct species.

We considered the set of the following 65 identified protein-coding genes in *R. americana*: *atp1, atp3, atp4, atp6, atp8, atp9, ccmA, ccmB, ccmC, ccmF, cob, cox1, cox11, cox2, cox3, nad1, nad10, nad11, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad8, nad9, rpl1, rpl10, rpl11, rpl14, rpl16, rpl18, rpl19, rpl2, rpl20, rpl27, rpl31, rpl32, rpl34, rpl5, rpl6, rpoA, rpoB, rpoC, rpoD, rps1, rps10, rps11, rps12, rps13, rps14, rps19, rps2, rps3, rps4, rps7, rps8, sdh2, sdh3, sdh4, secY, tatA, tatC, tufA*.

The NCBI Common Taxonomy Tree tool was used to construct a taxonomic tree for the species for which data was obtained. We then inferred the barcode of each branching point on the tree simply by applying the OR operator bitwise on each branch's barcodes. This protocol relies on two assumptions:

1. Mitochondrial gene loss is rare, so if two descending lineages of a common ancestor are observed to lack a gene, we assume that their common ancestor lacks it, as opposed to gene loss occurring convergently in both descendant lineages;

2. Mitochondrial gene *gain* is very rare [62], so if an ancestor has two descendants, one with and one without a gene, it is much more likely that that gene has been lost in one descendant than that it has been gained in the other.

Finally, we recorded the subset of edges on the tree where a change occurred between two connected mitochondrial genomes, from ancestor to descendant (Fig. 1). Our data $\mathcal{D}$ thus consists of n_D pairs of barcodes $\{a_k, d_k\}$, respectively ancestor and descendant in pair k.

We note that this approach allows us to account for the large sampling bias in recorded mitochondrial genomes. Many more vertebrate mtDNA sequences have been recorded than any other clade: however, mitochondrial gene sets are largely homogeneous across vertebrates. By only using instances where an independent change between parent and daughter has been observed, we ignore this oversampling and assign equal weights to each event of evolutionary change.

Inference of evolutionary dynamics

Mathematical background

We consider a state space consisting of all binary strings of length L, and a hypercubic transition network π describing the probability of a transition between two states. We assume that π is structured such as to ensure that a trajectory starting at 1^L terminates after L steps at 0^L, with each step involving a change of one locus in the current string from 1 to 0. The set of transition probabilities in π leading from any node are, by definition, constrained to sum to unity. We write $P(b|a;\pi)$ for the probability that a system initially at a will transition to b. The 'origin' state $O \equiv 1^L$ is where all evolutionary trajectories begin.

We will work in the picture of a hidden Markov model [74]. Here, the process of evolution in a single lineage corresponds to a trajectory across the hypercube, starting at 1^L and potentially ending at 0^L. Trajectories emit 'signals' at random with a characteristic rate. Each signal is simply the current state of the trajectory. There is thus a constant probability of emitting a signal at any state in a sampled trajectory. If some trajectories are more likely than others, and an ensemble of trajectories is simulated, more emitted signals will be expected from states in common trajectories than from states in rare trajectories.

The fundamental quantity of interest throughout is *the probability that randomly emitted signal(s) from a randomly sampled evolutionary trajectory match an observation.* The product of this probability over all independent observations (see below for discussion of independence) constitutes the likelihood of a given transition network given biological data.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Dynamics and likelihoods for evolutionary lineages

This subsection shows that the likelihood associated with data $\mathcal{D}$, a set of n_D observed transitions $\{a_i \to d_i\}$, is

$$\mathcal{L}(\pi|\mathcal{D}) = \prod_{i=1}^{n_D} P(d_i|a_i;\pi). \tag{1}$$

This result may hold intuitively for some readers; they are directed to the next subsection, where an efficient method for computing these probabilities is developed.

In the original application of phenotype landscape inference [46], the properties of observed species were assumed to be the product of convergent evolution, with each patterns of presence or absence the result of an independent evolutionary trajectory. Observations thus consisted of single states, where each observation was independent of all the other observations. The likelihood associated with an observation, as above, is the probability that a randomly emitted signal from a randomly sampled evolutionary trajectory matches the observation. To estimate this probability, random, independent evolutionary trajectories were simulated on π, and the proportion of states in each trajectory that were compatible with an observation was computed. This proportion was then proportional to the probability that a randomly emitted signal from a random trajectory matched the observation, with the constant of proportionality a multiplicative factor accounting for emission probabilities that was independent of model parameterisation and hence vanished in likelihood calculations.

The non-trivial taxonomic relationship between observations in the mitochondrial system means that this picture of independent trajectories leading to independent observations no longer applies. Instead we must consider the pattern of shared histories in the set of observations. For example, the joint probability of two observations known to come from the same mitochondrial lineage cannot be calculated as the joint probabilities of two independently sampled trajectories emitting signals compatible with those observations.

To give a concrete example, consider a system where we have observed the properties of two species: species A with `110` and species B with `100`. In the convergent evolution picture, species A and species B have both evolved independently from a common ancestor O with `111`, and the joint probability of these (independent) observations would simply be $P_{obs}(\texttt{111}\to\texttt{110})P_{obs}(\texttt{111}\to\texttt{100})$. However, if species B is a descendant of species A, the evolutionary processes are no longer independent: we know that both observations have been made in the same lineage. Hence, the joint probability is $P_{obs}(\texttt{111}\to\texttt{110})P_{obs}(\texttt{110}\to\texttt{100}|\texttt{111}\to\texttt{110})$. The important difference is that the non-convergent picture is 'serial', involving sequential observation probabilities contingent on previous steps, whereas the convergent picture is 'parallel', involving independent descents from the original common ancestor.

We have used $P_{obs}(\circ \to \circ)$ for the probability of *making an observation* as opposed to the probability $P(\circ|\circ;\pi)$ of undergoing a transition from one specific state to another. The two are different for two reasons. Firstly, to make an observation of a transition, we require a signal to be emitted at the initial and final state; there is a probability associated with each event, corresponding to the probability of emitting signals from the evolutionary process at a specific time. We write $P_{emission}$ to denote the probability of emitting signals at the required times. Secondly, the probability of observing a transition is generally proportional to the product of that transition probability and the probability $P_{reach}(a;\pi)$ of encountering its initial state. This captures the fact that, even if a transition to b is certain for a system at a, this transition will never be observed if the system never reaches a. In general, then, we have

$$P_{obs}(a \to b) = P_{emission}P_{reach}(a;\pi)P(b|a;\pi) \tag{2}$$

We will ignore the factors of $P_{emission}$ henceforth because if, as before, signals are emitted randomly and uniformly, independent of model parameterisation and the state of an evolutionary trajectory, $P_{emission}$ is a constant multiplicative factor that is a function of the data structure alone. This factor will then cancel when likelihood ratios are considered. Emission probabilities are discussed further, with examples, in 'Signal emission probabilities' below.

We are left with $P_{obs}(a \to b) = P(b|a;\pi)P_{reach}(a;\pi)$. In the convergent picture, $P_{reach}(a;\pi)$ vanishes as the initial state is always the origin state O where all trajectories start, and $P_{reach}(O) = 1$.

In the non-convergent picture, we have serial chains of observations, with the ancestral state in each observation being the descendant in another observation, except in the case of the origin state. We thus have a chained product of observation probabilities. For example, the observations $a \to b$, $b \to c$, and $c \to d$ in the same lineage give $P_{obs}(c \to d|b \to c)P_{obs}(b \to c|a \to b)P_{reach}(a;\pi)$. Each observation probability is contingent on having already reached the initial state through a different observation, and the initial state dependence thus vanishes for all but the most ancestral transition: for example, the term corresponding to an intermediate observation in the chain above, $P_{obs}(b \to c|a \to b) = P(c|b;\pi)P_{reach}(b|a \to b) = P(c|b;\pi) \times 1$. The origin state is encountered in every trajectory, as in the convergent case, and so the associated term is also unity.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Furthermore, we note that the initial state emission probability in Eqn. 2 is 1 if that state has already been emitted as the final state of a previous transition. Each observation thus only has one emission probability factor (which still cancels as discussed above).

The probabilities associated with evolutionary trajectories in parallel lineages are independent after their common ancestor and can straightforwardly be multiplied. It can then readily be seen that any combination of parallel and serial lineages reduces to a joint observation probability involving only the product of individual transition probabilities (an example is illustrated in Fig. 6A-B). Hence, for our inference purposes, it will suffice to consider the product of probabilities of observed transitions as the likelihood associated with a given dataset. We therefore have, neglecting multiplicative constants:

$$\mathcal{L}(\pi|\mathcal{D}) = \prod_{i=1}^{n_D} P(d_i|a_i;\pi) \tag{3}$$

An example of this calculation is given in Fig. 6C.

If all observed transitions $a_i \to d_i$ involved single changes, and hence single edges of π, each term in the likelihood function product could straightforwardly be read off from π. However, in general, a_i and d_i may differ at many positions, and many different trajectories may be used to transition between the two. We therefore require a way to calculate the probability of this transition, taking these different possible trajectories into account.

Efficiently estimating transition probabilities

We consider the problem of computing the probability that an evolutionary trajectory, beginning exactly at source s, will lead to the observation of target t, given transition matrix π. *Note that we here use t to denote a target state rather than time.* Transition probabilities $P(b|a;\pi)$ are determined by the edge weights of π; however, for clarity in the working below, we will not write this π dependence explicitly, and will adopt the symbol:

$$P(a \to b) \equiv P(b|a;\pi), \tag{4}$$

We are thus interested in the probabilities associated with all trajectories that lead from s to t. Labelling a trajectory consisting of N steps as $c = c^0, c^1, ..., c^N$, we have:

$$P(c) = \prod_{i=0}^{N-1} P(c^i \to c^{i+1}) \tag{5}$$

and so

$$P(s \to t) = \sum_c I(c,s,t) \prod_{i=0}^{N-1} P(c^i \to c^{i+1}), \tag{6}$$

where $I(c,s,t)$ is an indicator function returning 1 if trajectory c starts at s and ends at t and 0 otherwise.

Generally we expect this sum to be hard to perform through random sampling, as only a small number of all possible trajectories may pass through s and t. A very large number of randomly chosen trajectories will then need to be simulated to ensure that we characterise $P(s \to t)$.

We instead consider an approach where we constrain the trajectories we simulate to start at s and end at t, and account for the amount of bias we need to employ to do so.

Definition. A state r is t-compatible if $r_i = 1$ for all i for which $t_i = 1$.

Lemma 1. Each step on a trajectory that will eventually reach t must be t-compatible, as we cannot reacquire lost traits.

Define $T(c^i)$ as the set of t-compatible states that can be reached by acquisitions from point c^i. Consider two events: (a) a transition from c^i to c^{i+1}, written $c^i \to c^{i+1}$, and (b) a transition from c^i to any member of $T(c^i)$, written $c^i \to \circ \in T(c^i)$. From Bayes' Theorem:

$$P(c^i \to c^{i+1}) = \frac{P(c^i \to c^{i+1}|c^i \to \circ \in T(c^i))P(c^i \to \circ \in T(c^i))}{P(c^i \to \circ \in T(c^i)|c^i \to c^{i+1})}. \tag{7}$$

If $c^{i+1} \in T(c^i)$, the denominator $P(c^i \to \circ \in T(c^i)|c^i \to c^{i+1}) = 1$ and so

$$P(c^i \to c^{i+1}) = P(c^i \to c^{i+1}|c^i \to \circ \in T(c^i))P(c^i \to \circ \in T(c^i)). \tag{8}$$

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

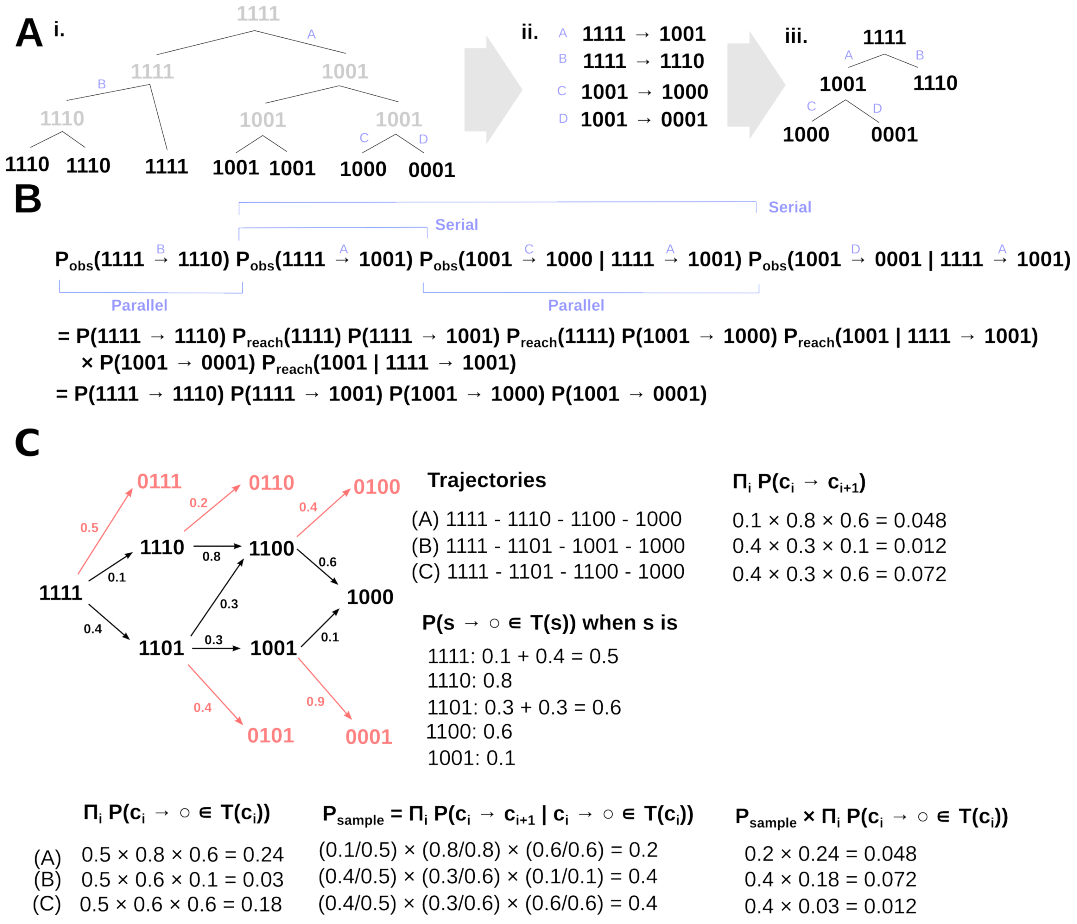

Figure 6: **Construction of source data and structure of likelihood calculations in HyperTraPS.** (A) (i) A set of observed strings (black) and taxonomic relationships (lines) is used to infer strings throughout a taxonomic tree (grey). This taxonomy then reveals changes in strings between ancestor and descendant (blue letters, (ii)). This set of changes forms an evolutionary tree (iii) describing the mitochondrial evolution occurring 'beneath' the speciation and adaptation of 'host' organisms, which may involve branching lineages of different structure from the species taxonomy. (B) The likelihood of this tree under a model of branching random walkers is then computed. For clarity we use the symbol $P(a \to b) \equiv P(b|a;\pi)$. The product of each observation probability is written down, including terms due to serial descent down a single lineage and parallel branching. The dependence of observation probability on encountering the correct initial state drops out, as the lineage structure means that every observed ancestral state is the result of another observed transition. A product of transition probabilities $P(a \to b) \equiv P(b|a;\pi)$ then constitutes the likelihood function. Note that some transitions may involve changing more than one property (for example, A, `1111` → `1001`), and may thus be accomplished through more than one trajectory. (C) HyperTraPS marginalisation and probabilities. An example transition network with edge weights is illustrated. Here, s = `1111` and t = `1000`. All strings with a `0` in the first position are thus not t-compatible. These states, and transitions to them, are coloured pink. t-compatible states and steps are marked in black. No other transitions are supported. Three trajectories A, B, C lead from s to t. Their individual probabilities $\prod_i P(c_i \to c_{i+1})$ are given at the top right. The other quantities involved in the main text are displayed for each trajectory, illustrating that the product of the sample probability of a trajectory $P_{sample}(c)$ and the product of the probability of making t-compatible steps at each state $\prod_i P(c^i \to \circ \in T(c^i))$ exactly matches the trajectory's individual probability.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

If $c^{i+1} \notin T(c^i)$, Eqn. 7 takes the form 0/0. However, this case involves a trajectory leading to a state that is not t-compatible, and thus cannot lead to t. Such a trajectory will therefore not contribute to the expression for $P(s \to t)$, which manifests mathematically as the observation that all subsequent steps j in such a trajectory will have $P(c^j \to \circ \in T(c^j)) = 0$. Hence the product of probabilities associated with this trajectory interpreted as a path between s and t is zero.

By Lemma 1, we have that a trajectory leading from s to t must have $c^{i+1} \in T(c^i)$ for all i. In this case, the associated probability is the product of Eqn. 8 for each step in the trajectory:

$$P(c) = \prod_i P(c^i \to c^{i+1}) \tag{9}$$

$$= \prod_i P(c^i \to c^{i+1} | c^i \to \circ \in T(c^i))) P(c^i \to \circ \in T(c^i)), \tag{10}$$

so the overall quantity of interest can be written

$$P(s \to t) = \sum_c \prod_i P(c^i \to c^{i+1}) \tag{11}$$

$$= \sum_c \prod_i P(c^i \to c^{i+1} | c^i \to \circ \in T(c^i)) P(c^i \to \circ \in T(c^i)). \tag{12}$$

The idea behind this recasting is to facilitate a sampling approach. Consider simulating a trajectory, starting at s and progressing according to the following rule. At each state c^i, identify the set of states $T(c^i)$ that may be reached by one acquisition that are t-compatible. Compute $p^j = P(c^i \to c^j) / \sum_{r \in T(c^i)} P(c^i \to r)$ for each member c^j of this set, where the sum over r is taken over all members of the set. Choose the next step according to the probabilities p^j. This rule enforces t-compatibility in each step of the trajectory, thus forcing every trajectory to transition between s and t, while retaining the correct relative weighting of steps between states. p^j, the probability of transitioning to state c^j given that a transition is made to a t-compatible state, is identically $P(c^i \to c^{i+1} | c^i \to \circ \in T(c^i))$ when $j = i + 1$. The factor of $\prod_i P(c^i \to c^{i+1} | c^i \to \circ \in T(c^i))$ is thus exactly the probability with which trajectory c will appear when simulations are performed over the set of trajectories constrained to begin at s and end at t. We can thus write

$$P_{sample}(c) = \prod_i P(c^i \to c^{i+1} | c^i \to \circ \in T(c^i)), \tag{13}$$

where $P_{sample}(c)$ is understood to mean the probability of simulating trajectory c given the above protocol. If we then adopt this sampling scheme and record the average value of $\prod_i P(c^i \to \circ \in T(c^i))$ for each sample, we will obtain an estimate of the sum in Eqn. 12:

$$\hat{P}(s \to t) = \sum_{sampled\, c} P_{sample}(c) \prod_i P(c^i \to \circ \in T(c^i)) \tag{14}$$

$$= \left\langle \prod_i P(c^i \to \circ \in T(c^i)) \right\rangle_{sampled\, c} \tag{15}$$

The advantage of this sampling approach is that we can simulate trajectories guaranteed to start at s and end at t, thus avoiding wasting computational time on trajectories that do not contribute to the overall sum. Convergence of $\hat{P}(s \to t)$ is therefore expected to proceed much more quickly than a naïve sampling approach involving unconstrained trajectories. An illustration of the quantities above for a specific transition network is shown in Fig. 6C.

The HyperTraPS Algorithm

Here we describe our algorithm for hypercubic transition path sampling. We are not aware of this algorithm having been previously published; if this is true, we propose the name 'HyperTraPS', both standing for **hyper**cubic **tra**nsition **p**ath **s**ampling and referring to the act of forcing trajectories towards specific points on a hypercube.

Algorithm 1. *Hypercubic transition path sampling (HyperTraPS)*

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1. Initialise a set of N_h trajectories at s.
2. For each trajectory i in the set of N_h:
   (a) Compute the probability of making a move to a t-compatible next step (for the first step, all trajectories are at the same point and the probability for each is thus the same); record this probability as α'_i.
   (b) If current state is s, set $\alpha_i = \alpha'_i$, otherwise set $\alpha_i \to \alpha_i \alpha'_i$.
   (c) Select one of the available t-compatible steps according to their relative weight. Update trajectory i by making this move.
3. End for each.
4. If current state (in all trajectories) is t go to 5, otherwise go to 2.
5. $\hat{P}(s \to t) = N_h^{-1} \sum_i \alpha_i$.

In this algorithm, α is a vector, with each of N_h elements progressively recording the product of probabilities $\prod_i P(c^i \to \circ \in T(c^i))$ of transitioning to any t-compatible state, where the product is taken over all steps so far performed in the corresponding trajectory. Each trajectory is simulated as above, by choosing a t-compatible transition at each state according to its relative weight. When t has been reached, the average $\prod_i P(c^i \to \circ \in T(c^i))$ is computed over all sampled trajectories.

N_h, the number of sampled trajectories, is a parameter of the algorithm. Lower numbers will be computationally cheaper but will give a poorer sampling of possible trajectories and thus a less accurate estimate $\hat{P}(s \to t)$.

Inferring transition matrices

Experimental observations of mitochondrial genomes constitute a dataset $\mathcal{D}$, which consists of a set of paired barcodes a_k, d_k, respectively the kth ancestral and descendant barcodes. We computed the likelihood $\mathcal{L}(\pi|\mathcal{D})$ associated with a trial transition matrix π by sequentually using each a_k, d_k pair as the s, t pair in the HyperTraPS algorithm, and multiplying the likelihoods associated with each pair. For computational convenience, log-likelihoods $l = \log \mathcal{L}$ were used.

These log-likelihoods were then used in a Bayesian MCMC framework, with a new trial transition matrix π' being produced from a current transition matrix π by applying Gaussian-distributed perturbations with standard deviation $\sigma = 0.25$ to each element of π. An initial parameterisation of π enforcing uniform gene loss probabilities with no evolutionary contingency was employed. $N_h = 200$ HyperTraPS trajectories were used to estimate likelihoods. 10^6 MCMC iterations were used with a burn-in period of 2×10^5 iterations, with posterior samples taken every 10^3 iterations. N_h and σ were chosen through preliminary investigation to lead to good chain mixing and convergence. $N_c = 10$ repeats were performed with different random number seeds to check convergence and facilitate clean statistics. The posterior distributions on network parameters were visualised as described in the Main Text.

Validation

To test the performance of the HyperTraPS algorithm, we first considered the simplest possible case, where all π are equal (and thus all possible transitions from every point equally likely), where the start node $s = 0^L$ is the node corresponding to an all-zero barcode, and where t is a single, specific target. The probability of encountering t is then the probability of sequentially acquiring the n_t traits of t without acquiring any others. This probability is $(n_t/L) \times ((n_t - 1)/(L - 1)) \times ...$, hence:

$$P(t) = \prod_{n_r=0}^{n_t} \frac{n_t - n_r}{l - n_r} \tag{16}$$

which is straightforwardly enumerable. We verified for a simple $L = 10$ systems that the HyperTraPS observation frequency exactly matched this analytic result for a range of random targets (Fig. 7A).

Next, we relaxed the restriction on π to consider more general transition matrices that lacked a straightforward analytic result for the observation probability of a given target. Once more, we considered the probability of observing a set of randomly chosen target strings, and compared the HyperTraPS results with those estimated from sampling many explicitly simulated random trajectories on the given hypercube. We fixed π at $L = 10$ and explored the correlation between the HyperTraPS and simulated results, which is excellent over many randomly chosen targets (Fig. 7B).

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

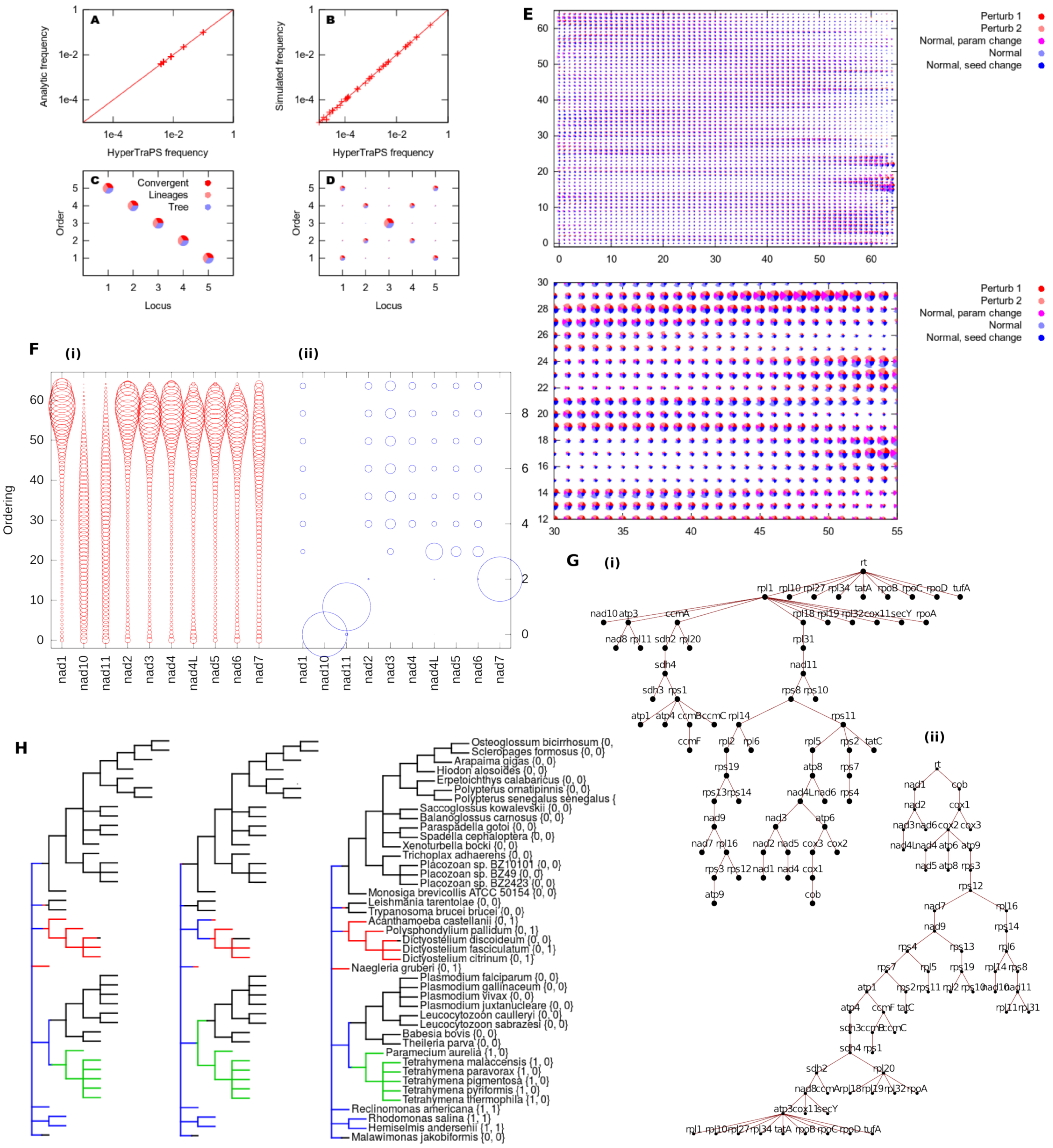

Figure 7: **Validation of HyperTraPS algorithm and comparison to other approaches.** (A) Comparison of HyperTraPS observation frequency and analytic observation frequency for simple π and random single specific targets. (B) Comparison of HyperTraPS observation frequency and simulated observation frequency for non-trivial π and random targets. Departures at low frequencies are due to the finite sampling allowed by the simulation approach (2×10^6 trajectories simulated for each target). (C-D) Reconstructed evolutionary dynamics using HyperTraPS and Bayesian inference, for (C) a simple underlying hypercube supporting a single evolutionary path and (D) a more complicated hypercube supporting two distinct pathways. A range of taxonomic relationships (covergent, linear, branching tree) between observations were considered in each case (see text). (E) Robustness of results with respect to perturbations in data gathering and simulation protocol. The posterior probabilities, represented by arc radii, associated with gene loss ordering for short illustrative simulations run with several different protocols. The 'perturb' label involves two randomly perturbed datasets as described in the text, testing robustness with respect to specific details of the phylogenetic and mtDNA structure reconstructions. The 'normal' involves the unperturbed dataset with default parameters; results using different simulation parameters ($N_h = 100$, $\sigma = 0.1$) and a different random number seed are all shown. (top) Full 65×65 set of posteriors; (bottom) zoomed view on a particular region. Gene ordering (horizontal) is arbitrary in these plots and gene labels are omitted for clarity. All results are very comparable and do not lead to substantial changes in our general results. (F-H) Other related methods for trait evolution over related observations. (F) 'OrderMutation' inference of gene loss orderings given a reduced set of *nad[X]* genes and a reduced set of observations. (i) Shows the posterior distributions on loss orderings from our approach; (ii) shows the mean loss orderings from OrderMutation over 20 bootstrap resamples. (G) 'Oncotrees' inference of maximum likelihood trees of trait relationships. (i) Loss treated as the 'acquisition' of a loss event (root is 000...). (ii) Loss treated as explicit losses, inverting the normal action of oncotrees (root is 111...). In both cases, *cob*, *cox[X]* and many *nad[X]* genes are inferred to be lost late, and rare genes (for example, *secY*) are inferred to be lost early. Some structure corresponding to different complexes is visible: some *nad[X]* genes cluster together, as do some ribosomal genes. (H) 'Simmap' stochastic mapping of trait evolution on a reduced phylogeny. Three instances of stochastic maps of the evolution of { *nad10*, *nad11* } on a reduced phylogeny. Colours give states: $\{1,1\}$ blue; $\{1,0\}$ green; $\{0,1\}$ red; $\{0,0,\}$ black.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Finally, we tested the HyperTraPS algorithm in a controlled inferential setting. We constructed simple $L = 5$ artificial datasets consisting of samples from random walks on transition hypercubes with known transition probabilities, and attempted to infer these known underlying dynamics from the artifical data with HyperTraPS. The two transition hypercubes we used represented a simple case where a single evolutionary pathway is supported, and a more complicated case where two distinct pathways exist, and the trajectory experienced by a lineage is contigent on which of two equiprobable first steps is taken. For each transition hypercube, we constructed artificial data to cover a range of phylogenetic structures: totally convergent evolution (where no observed species is related); evolution involving the minimal number of possible lineages (one for a single pathway, two for two competing pathways); and a branching-tree phylogeny linking observed species in a more complicated arrangement. We thus covered a range of underlying evolutionary dynamics and a range of taxonomic connections between observations. In these cases, we used the reduced parameterisation of π described in the Methods section (using $O(L^2)$ rather than $O(2^L)$ parameters while maintaining estimates of independent and contingent loss probabilities). Fig. 7C-D illustrate the excellent reconstruction of the dynamics supported by the underlying transition matrix in each case.

To confirm that our results were not dependent on specific details of the reconstructed phylogenetic tree from the NCBI Taxonomy Tool, or the inference of gene loss events, we applied random perturbations (changing presence/absence properties of individual genes with probability 0.05) to the set of ancestor-descendant pairs that arose from our data analysis. These random perturbations model the effects of changes to the phylogenetic structure and inferred ancestral mtDNA structures across our dataset. In parallel, to confirm that simulation runs were not becoming trapped and that chains were successfully mixing to converge on a true posterior, we compared the results from several different simulation protocols, involving different random number seeds and step sizes (see Fig. 7E). The agreement between the posteriors in all these cases confirms the convergence of the algorithm and its robustness to perturbations in the source data.

Existing methods

In the Main Text we discuss the similarities and differences between our approach, other methods from the systematics literature for exploring trait evolution on a phylogeny, and other methods (largely) from the cancer progression literature for inferring the ordering of a set of discrete, irreversible transitions. Fig. 7F-H illustrates the output of several of the methods most closely aligned with our approach.

For 'Oncotrees' [47], the full set of 2015 observations was treated as input. For 'OrderMutation' [48], we use only the set of $n = 74$ unique genomes as input, and a reduced set of $L = 10$ genes from this set, namely { *nad1*, *nad10*, *nad11*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7* }, and bootstrap over 20 resamples, each using 10 simulations, to estimate mean orderings. For 'Simmap' [75, 41], a reduced phylogeny consisting of a subset of species was used for clarity (and to better illustrate diverse evolutionary trajectories, highlighting some unicellular branches); all branch lengths were set to 1. The genes { *nad10*, *nad11* } were chosen to reflect diverse behaviour in this reduced phylogeny. Transitions that involved the acquisition of genes were prohibited by using a lower-diagonal transition matrix model. The output of 'Simmap' assigned equal rates (0.20) to each remaining possible transition between states, implying no asymmetry in transitions for this (reduced) example.

Gene data acquisition and curation

Table 1 summarises the genetic and physical properties of genes that we consider in the model selection process.

Compilation

The **length*** (in bases) and **GC content** (average number of G bases and C bases per codon) of genes are taken straightforwardly from a sequence. The **GC skew*** was computed as $(G - C)/(G + C)$, where G is the number of G bases and C the number of C bases [12], from the gene sequence. The **strand*** upon which a given gene was encoded was represented by a 0 (light strand) or 1 (heavy strand) according to GOBASE's entry.

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. [76]). **Amine group pK$_a$**, **carboxyl group pK$_a$**, and **molecular weight*** values were calculated using this compilation (original data from [66]).

Glucose energy costs* were computed using the $A_{glucose}$ metric, based on the absolute nutrient cost required for amino acid biosynthesis, from Ref. [67]. **Craig-Weber energy costs***, estimating the number of high-energy

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Property	Code	Derivation	Related hypothesis	Pearson's r between *R. americana* and average
Length	Leng	Simple length L of a gene in bases	Longer genes or larger protein products are harder to transfer	0.97
Molecular weight	M_We	Summed relative molar mass of amino acids in protein product [66]	Heavier protein products are harder to import	0.97
Hydrophobicity	Hydr	Hydrophobicity of amino acids in protein product, normalised at pH 7 with glycine set to 0 and maximum 100 [76] (hydrophobicity index measure I_Hy also used)	Hydrophobic proteins are harder to import to the mitochondrion from the cytosol and so are retained locally [20]	0.55
Energetic cost	Aglu	Estimated (unitless) amino acid synthesis energy burden from Ref. [67] (alternative energy measure CWEn from Ref. [77], in number of ATP molecules, also used)	Energy required to produce a protein may influence its production location	0.97
pK_a	pKa1	Averaged carboxyl pK_a values for protein product [66] (amino pK_a pKa2 also used)	Possible signature of different chemical behaviours in mitochondrial matrix and nucleus or cytosol	0.20
GC content	GC_c	Number of guanine or cytosine bases per codon in a gene, $(G + C)/3L$	Thermodynamic stability of DNA [27] or transcripts affects retention	0.67
GC skew	GC_s	$(G - C)/(G + C)$	Possible role for asymmetric mutation pressure in mtDNA [26, 12]	0.39
Mutational robustness	Robu	Proportion of non-synonymous point mutations	Genes more susceptible to mutational damage are safer in the nucleus [13]	0.64
Universality	UniN	Index of number of codons whose interpretation varies across life (alternative UniI index also considered)	Genes incompatible with nuclear genetic code are harder to transfer [14]	0.51
Strand	Stra	Which mtDNA strand gene is encoded on (0, light; 1, heavy)	One strand preferentially retained over another	0.54

Table 1: **Gene properties and hypotheses compared in Bayesian model selection.** The different properties of genes and gene products used to explore hypotheses regarding mitochondrial gene retention. ETC, electron transport chain. Illustrative Pearson's r correlation coefficient computed between the values of genes in *R. americana* and the values of genes averaged across all available eukaryotic genomes.

Amino-acid-code-3	Amino-acid-code-1	Hydrophobicity	Hydrophobicity-class	Molecular-weight	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 2: **Amino acid properties used in model selection.** Numerical values of the properties described in the text. See text for sources.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Codon	Amino-acid-code-3	Amino-acid-code-1	Non-universal	Non-universal-index
CUA	Leu	L	1	1
CUC	Leu	L	1	1
CUG	Leu	L	1	1
CUU	Leu	L	1	1
UAA	Stop	X	1	1
UCA	Ser	S	1	1
UAG	Stop	X	- (1)	- (2)
AAA	Lys	K	1	3
AUA	Ile	I	1	5
AGA	Arg	R	1	7
AGG	Arg	R	1	7
UGA	Trp	W	1	9

Table 3: **Non-universal codons.** The set of codons for which mitochondrial interpretation has been observed to differ in different taxa. The non-universality index counts the number of taxonomic cases for which such departure has been observed.

phosphate bonds and reducing hydrogen atoms required from the cellular energy pool to produce an amino acid, were taken from Ref. [77]. These biochemical properties are summarised in Table 2.

Robustness was assigned by counting the number of point mutations that would lead to nonsynonymous changes in the gene product. The **universality** and **universality index** of a gene were defined by considering the number of codons in the gene that were subject to different interpretations throughout the Tree of Life. These codons were identified using the NCBI compilation at `http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi`. The non-universality of a codon was set to 1 if its interpretation was ever different from the universal genetic code, and 0 otherwise; the non-universality index of a codon was equal to the number of different taxa in which it has been observed to differ (see Table 3).

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.

The **assembly energy** of genes was quantified using the PDBePISA tool [70, 71] (`http://www.ebi.ac.uk/pdbe/pisa/`) to analyse the energetic interactions of the subunits in solved structures of electron transport chain complexes in the PDB [72] (specifically, Complex II (PDB 2h88, *Gallus gallus*), Complex III (PDB 3cxh, *Saccharomyces cerevisiae*), and Complex IV (PDB 1oco, *Bos taurus*)). The chains corresponding to products of genes in our reference set were identified, and for each chain, the total energy of all interfaces with other chains was recorded as its assembly energy.

We also explored the link between **gene expression levels** and evolutionary history of mitochondrial genes. The level of gene expression in animals has been postulated to affect the rate of sequence evolution of mitochondrial genes [28], possibly due to the deleterious effects of misfolding abundant proteins. To explore this link we required gene expression data of mtDNA genes, preferably in species with large numbers of mtDNA genes. We obtained RNA-seq data from *Phoenix dactylifera L.* (date palm) [69] and *Lolium perenne L.* (perennial ryegrass) [68], both containing ~ 30 genes. We use data from mitochondria in *Lolium* and data from green leaves in *Phoenix*; other tissue types in the latter did not qualitatively change our findings.

Fig. 8D shows the links between gene expression levels in these species and inferred gene retention ordering and GC content in *R. americana* and across species. In *Lolium*, little correlation is present between gene expression and retention, or between expression and GC content. In *Phoenix*, we observe a moderate correlation between expression levels and gene retention. Stronger correlations are present between expression levels and GC content, particularly GC content from *R. americana*. This overall pattern of links is not compatible with a universal link between gene expression and mtDNA gene retention. However, it is compatible with a picture where a link between expression levels and gene retention is observed in some species where expression correlates with GC content (*Phoenix*), but not in other species (*Lolium*).

Correlations and consistency

All properties except assembly energy were computed both using *R. americana* as a reference genome and by taking the average value across all organisms in which the given gene was present. Most properties (with GC skew a notable

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

exception) showed at least reasonable correlations between these two approaches (Table 1), suggesting that these very coarse-grained mitochondrial genetic properties do not vary beyond recognition between organisms. GC skew displayed dramatic differences between different organisms, suggesting that other evolutionary pressures may act on this more detailed genetic feature [26, 12] (see Discussion).

In Fig. 8D-E we illustrate the correlations between the different features used in this study. The strong correlations between the following physically similar pairs of features motivated the removal of one of the pair from the set of features explored in the main text: Hydr and I_Hy; pKa1 and pKa2; Aglu and CWEn; UniI and UniN.

We particularly focus on the connections between those features that we identify as key influences on mitochondrial gene loss: GC content, hydrophobicity, and energetic centrality. As discussed in the Main Text, these three features may intuitively be thought of as biologically connected: GC-rich codons encode hydrophobic amino acids, and hydrophobic peptide chains occupy central positions in complexes and within membranes. The weak and inverse ($r = -0.34$) correlation between GC content and hydrophobicity in the genes we consider (Fig. 8F) immediately suggests that this link may not represent the full story. Fig. 8F pursues possible links between these variables further, showing that none of these features has substantial predictive power for the others over the set of genes that we consider.

Model selection

The values of gene properties across the set of L genes under consideration were collected and normalised to form a matrix g_{ij}, where g_{ij} is the normalised value of property j for gene i. This normalisation was performed by simply dividing each value by the maximum observed value for that property over all genes, thus ensuring that $g_{ij} \in [0, 1]$.

Model selection proceeded by using a trial vector α to represent the coefficients of each of N features under consideration in a trial model.. The sum

$$y_i = \sum_j \alpha_j g_{ij}, \tag{17}$$

where g_{ij} is the value of property j for gene i, was computed. The L genes under consideration were then ordered by ascending values of y_i, yielding a vector of ranks ρ, where ρ_i is the rank of gene i in the ordered list. Given the inferred posterior ordering P_{ij} derived from HyperTraPS, we compute the likelihood associated with model α as $\mathcal{L} = \prod_i P_{i\,\rho_i}$, the joint probability of each gene being lost in the order predicted by the model.

Bayesian MCMC was used to perform model selection given this likelihood definition. Two different prior protocols were used.

Uniform priors. Each step, a uniform random number between 0 and 2^N was chosen. The binary representation of this number determined the features to be included in the model: if the ith bit in the binary representation of the number was zero, the ith element of α was set to zero. This procedure thus assigned a uniform prior distribution over each of 2^N possible model structures.

Exponential priors. An exponentially-distributed random number with mean 2 was produced. This number gave the number of non-zero features to include in a trial model. This mean was chosen to allow a reasonable probability of choosing a full set of features, while also strongly favouring more parsimonious models.

Non-zero elements of α were assigned uniformly distributed random values on $[-\lambda, \lambda]$, where $\lambda = 5 \times 10^3$ was chosen to exceed the range of model parameterisations found through a preliminary bootstrapping investigation. MCMC was used to sample from the posterior distribution of model structures and model parameterisations. 10^9 samples were performed, chosen to give sufficient coverage to each of the $2^{13} \simeq 10^5$ possible models. Figs. 8A-C demonstrate the consistent favouring of GC content and hydrophobicity across this set of protocols (different priors and different sizes of feature set).

The Main Text shows the correlation between inferred gene loss ordering and the predictions of a statistically-supported model involving GC content and hydrophobicity. Fig. 8G demonstrates that these predictions also correlate well with direct biological observations – specifically, the number of distinct mtDNA structures that contain a given gene. Thus, genes with high GC content and high hydrophobicity are observed in many different mtDNA structures; those with low GC and hydrophobicity are observed much less frequently.

In reporting a p-value associated with the link between a given model and inferred retention properties, the fact that we have selected one of many models and used a fit parameterisation must be taken into account. However, even given the substantial multiple hypothesis correction required when dealing with 2^{10} different model structures, the low p-value of 2×10^{-11} provides a compelling suggestion that the link between GC content, hydrophobicity, and retention is not coincidental.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

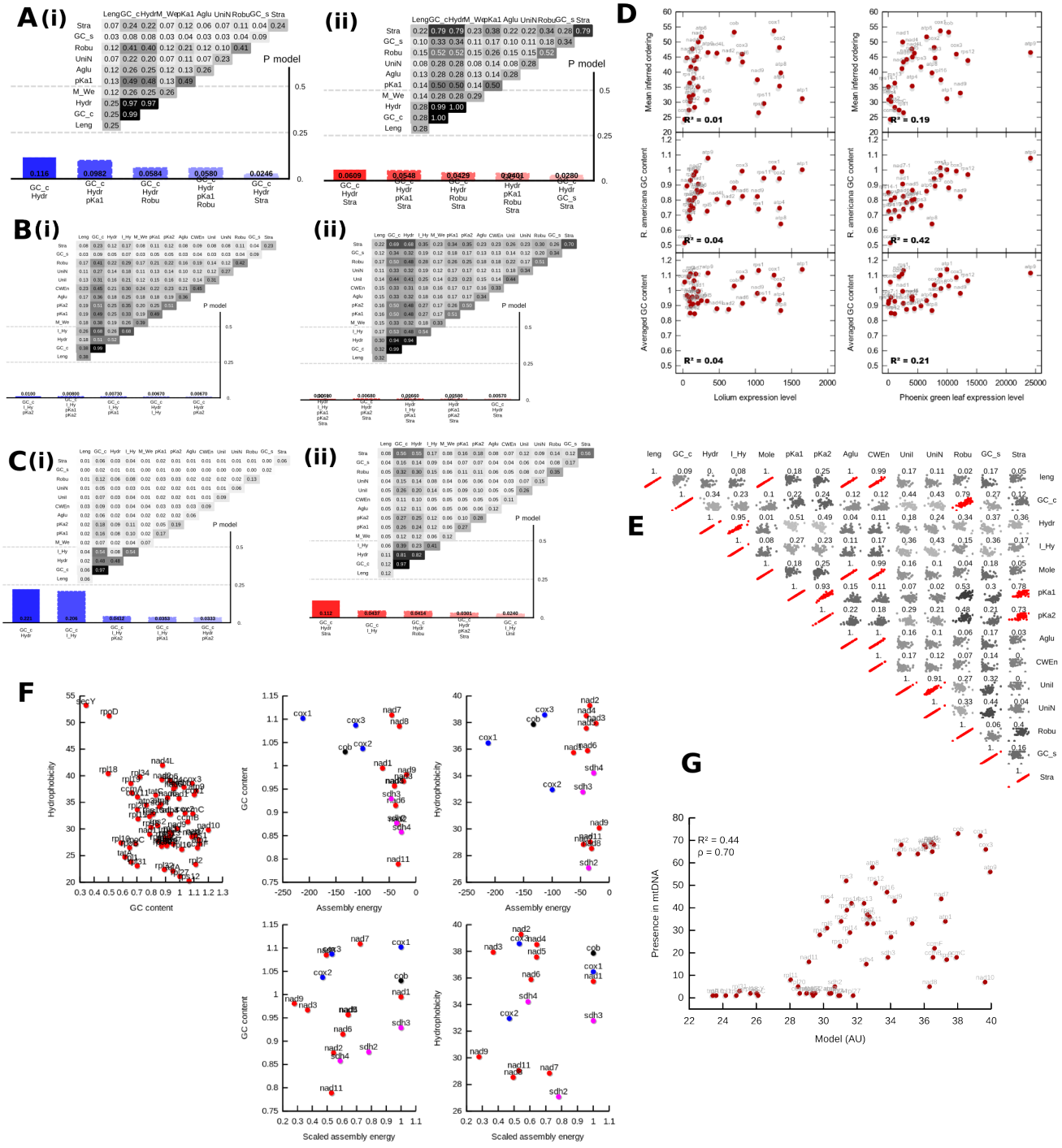

Figure 8: **Model selection under different protocols, and correlations and links between gene features.** (A) Uniform priors on model structure with default feature set; (B) uniform priors on model structure with expanded feature set; (C) exponential priors on model structure with expanded feature set. As before, columns show the support for the most supported model structures, and matrices show the proportion of inferred models in which a given feature (diagonal elements) or pair of features (off-diagonal elements) occurs; (i) show model selection based on *R. americana* features, (ii) based on taxa-average features. (D) Correlations between gene expression levels, inferred loss ordering, and GC content in two plant species (see text). (E) Correlations between different model selection features. Scatter plots (axes omitted) and Pearson's r for correlations between pairs of features used in model selection, using the values derived from cross-species averaging. (F) Correlations between features identified as linked to gene loss. Relationships between GC content, hydrophobicity, and assembly energy for all genes considered (GC vs hydrophobicity, as in the corresponding element of Fig. 8E) and ETC subunits (assembly energy). Scaled assembly energy gives each subunit's energy scaled by the highest-magnitude energy within that protein complex. Little correlation exists between any variables; the *cox[X]* and *sdh[X]* genes display a moderate link between GC content and scaled assembly energy (bottom centre) but the null hypothesis of no correlation cannot be discarded ($p > 0.05$ in both cases). GC content and hydrophobicity are unitless; unscaled assembly energy has units of kJ mol^{-1}. (G) Correlation between fitted model and mtDNA observation patterns. As in the Main Text, a model involving GC content and hydrophobicity scores for each gene was constructed and parameterised to fit the inferred mean of gene loss ordering. This plot shows values for each gene under that parameterised model, against the number of distinct mtDNA structures observed in our dataset that contain that gene. GC content and hydrophobicity thus exhibit a strong link to observed gene occurrences, as well as loss ordering (Main Text).

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

MtDNA-wide GC content across taxa

Our analysis provides strong support for genes with high GC content *relative to other genes in the same organism* being preferentially retained in mtDNA. Patterns of GC content vary dramatically between species [27]. Notably, at a sequence level, we have found that protein-coding genes in mtDNA can in some species display a bias *against* GC content, in the sense that GC-poor codons are used more often than GC-rich codons to encode a given amino acid (see below). However, the observation that GC-rich codons are less frequent in the mtDNA of some species does not necessarily conflict with our finding that genes with *relatively* high GC content are preferentially retained. Asymmetric mutational pressure generally acts at the sequence level to reduce GC content and enrich GC-poor codons in organellar genomes [26]. We propose that selective pressure at the structural level generally acts to retain genes with higher GC contents, against a background reduction in GC content across the mtDNA genome from asymmetric mutation. Our results thus suggest a tension between an entropic mutational drive at the sequence level (decreasing GC content in codons) and a selective drive at the genomic level (retaining genes with higher GC content).

To investigate trends in GC content throughout the mitochondrial genomes of different species, we first computed a 'null model' describing the expected pattern of codon usage if every codon encoding a given amino acid was used in the genome with equal probability (no favouring of one codon over another). We then recorded the actual patterns of codon usage in individual mtDNA genomes from sequence data, and compared the two. In Fig. 9A we illustrate the observation that *R. americana* displays an observable bias against those codons with high GC content, preferentially using codons with lower GC content to encode a given amino acid. No such bias exists in *H. sapiens*. This variability in GC codon bias is observed across taxa.

The observation of species displaying a genome-wide bias against high GC codons are not incompatible with our findings that genes with high GC content are preferentially retained in mtDNA. A genome-wide pressure against high GC codons can exist independently of an inter-gene favouring for high GC content. Thus, even if an organism's environment or biochemistry strongly favours low GC codons throughout the full mtDNA sequence, our statement that genes with a *relatively* high GC content are retained is unaffected.

One reason for the observed link between GC content and retention could be an indirect relationship through different levels of structural conservation between mtDNA genes. In this picture, asymmetric mutation pressure drives a reduction in GC content, but structural constraints are stronger in highly conserved genes, meaning that these genes retain high GC content. If highly conserved genes are preferentially retained in mtDNA (perhaps due to their structural importance in complex assembly), highly retained genes will passively display higher GC content.

To test this hypothesis we examined the GC content at different positions within the (less synonymous) first and second positions and (more synonymous) third positions in each codon in mtDNA genes, reasoning that structural conservation on a background of asymmetric mutation pressure would lead to higher GC content at nonsynonymous sits. In *R. americana*, we do observe (Fig. 9B) a notably lower GC content in synonymous loci, suggesting that (assuming a uniform initial condition) nonsynonymous loci are under higher selective pressure to retain their GC content. However, this behaviour is less clear in taxa-averaged gene data and vanishes for *H. sapiens* (Fig. 9B), suggesting that across taxa the link between GC content and simple structural conservation is less pronounced than in *R. americana*.

We sought to determine how much a potential link between structural conservation and mtDNA gene retention could explain the observed signal associated with GC content. To explore this question, we included both total GC content and position-specific GC content in our model selection process. If gene retention propensity is solely determined by degree of structural conservation, we would expect GC content at nonsynonymous positions (interpreted as a proxy for conservation) to be favoured over total GC content in model selection. Instead, we see an even combination of total GC content and nonsynonymous GC content in *R. americana*, and a pronounced favouring of total GC content in taxa-averaged gene properties (Fig. 9C). We conclude that while a link between structural conservation and gene retention can explain some of the signal associated with GC content, it cannot be solely responsible for the appearance of this signal across species.

Notes and extensions

Signal emission probabilities

We have assumed throughout that the probability associated with emitting signals that correspond to observations provides only a multiplicative constant factor independent of a specific model parameterisation. Here we discuss and illustrate the details of these emission probabilities.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

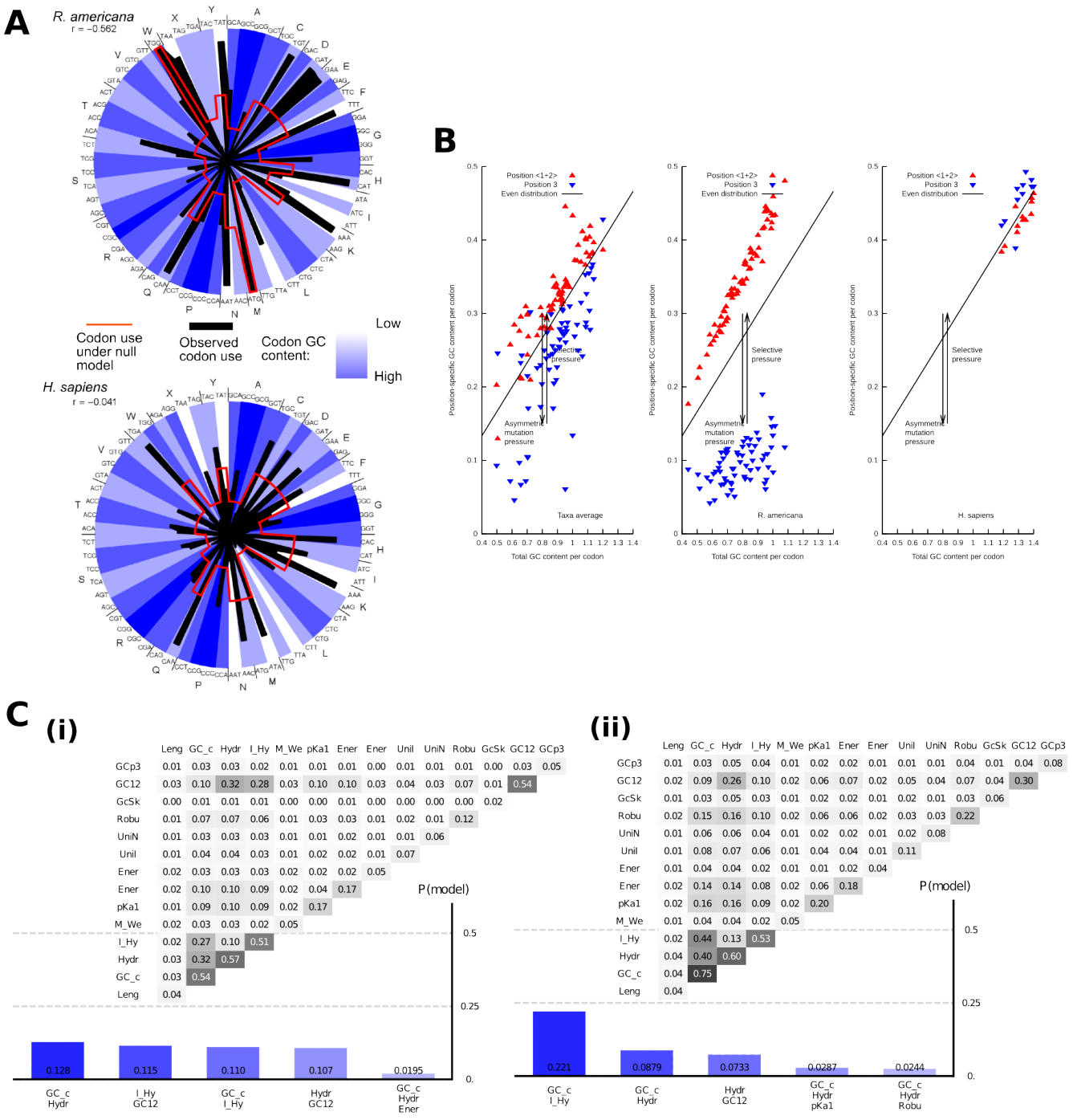

Figure 9: **GC content within and across species.** (A) Codon use and GC content. The black polygon gives the proportion usage of each codon (segments) that can encode each amino acid (groups of segments) in protein-coding genes in (top) *R. americana* and (bottom) *H. sapiens* mtDNA. Codon segments are shaded according to GC content (blue highest, white lowest) and a null model where each codon for a given amino acid is used equally is shown in red. In *R. americana*, GC-rich codons are disfavoured: the black usage polygon often falls below the red null model polygon in GC-rich (blue) segments. This disfavouring is quantified by considering Pearson's r between the amount of disfavouring (the ratio observed codon usage to codon usage under the null model) and GC content (top left). *H. sapiens* displays little correlation between codon use and GC content. (B) GC content at more synonymous (position 3) and less synonymous (average of position 1 and 2) positions in codons, for different mtDNA genes (datapoints) in *R. americana*, *H. sapiens*, and averaged across taxa. In *R. americana*, nonsynonymous positions display markedly lower GC content, suggesting a role for structural conservation; this link is less clear in the taxa-averaged and *H. sapiens* data. (C) Model selection using nonsynonymous (GC12) and synonymous (GCp3) GC content as well as total GC content (GC_c) in (i) *R. americana* and (ii) taxa-averaged data. Total GC content still plats a dominant role in the most explanatory models.

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

In the first application of phenotype landscape inference [46], observations were independent (due to convergent evolution) and consisted of individual signals, sometimes containing uncertain data. Recall that in our mathematical model, signals are emitted uniformly from an evolving system with a given probability $P_{emission}$ at each state. The probability of making a specific observation is then the probability of the system being in a state compatible with the observation when a signal is emitted. As $P_{emission}$ was a constant of the model, this observation probability is straightforwardly proportional to the number of states in sampled trajectories that match that observation.

We now discuss the emission of signals down a lineage in more detail. Consider the case where we have n observations known to arise from the same lineage, and that these observations are time-ordered. What is the probability with which a system, randomly emitting signals, emits a set of signals at the specific times matching an observation?

To compute this probability, we first assume that the system will emit exactly n signals over a trajectory of length L, and that the emission of these signals is random and uniform over the trajectory. The probability of emitting exactly n states is a function of the assumed emission rate λ alone, independent of the data and transition network. We write $P_\lambda(n)$ for this probability.

Assuming that n signals are emitted, and thus picking up a factor $P_\lambda(n)$ in our overall probability, we can represent the times of signal emission as n independent random variables taking values between 0 and $L-1$, where each value corresponds to the number of steps from the origin state that have occurred when a signal is emitted. We are interested in the probability with which, when ordered, the set of these timings matches the observation pattern we require. We write this pattern as $\mathbf{s} = \{s_1, ..., s_n\}$ where $s_1 \leq s_2 \leq ... \leq s_n$, with each element again corresponding to the number of steps from the origin state at which an observation is made.

Denote by m the number of distinct values taken by the elements of $\mathbf{s}$, and by $c_1, c_2, ..., c_m$ the number of times each of these values occurs. Thus, $\mathbf{s} = \{0, 0, 1, 2, 3, 3\}$ has $m = 4$ and $\mathbf{c} = \{2, 1, 1, 2\}$. The number of emission patterns that lead to the observation of $\mathbf{c}$ upon time ordering is the number of ways that this set can be ordered, taking identical values into account, which is $\frac{n!}{\prod_{i=1}^{m} c_i!}$. The probability of observing a given $\mathbf{s}$ is then this degeneracy multiplied by the probability with which each value is sampled:

$$P(\mathbf{s}) = P_\lambda(n) \frac{n!}{\prod_{i=1}^{m} c_i!} \prod_{i=1}^{m} \left(\frac{1}{L}\right)^{c_i} \tag{18}$$

$$= P_\lambda(n) \frac{n!}{\prod_{i=1}^{m} c_i!} \left(\frac{1}{L}\right)^{n} \tag{19}$$

When each observation is known to have occurred at specific states, the ordering is fixed, and hence $P(\mathbf{s})$ is a constant. For example, the observations $\texttt{111} \to \texttt{101}$, $\texttt{101} \to \texttt{000}$ correspond to $\mathbf{s} = \{0, 1, 3\}$. We then have

$$P_{obs}(\texttt{111} \to \texttt{101 and 101} \to \texttt{000}) = P_{obs}(\texttt{111} \to \texttt{101}) P_{obs}(\texttt{101} \to \texttt{000} | \texttt{111} \to \texttt{101}) \tag{20}$$

$$= P_{emission} P_{reach}(\texttt{111}) P(\texttt{101} | \texttt{111}; \pi) P_{reach}(\texttt{101} | \texttt{111} \to \texttt{101}) P(\texttt{000} | \texttt{101}; \pi) \tag{21}$$

$$= P_\lambda(3) P(\mathbf{s} = \{0, 1, 3\}) P(\texttt{101} | \texttt{111}; \pi) P(\texttt{000} | \texttt{111}; \pi) \tag{22}$$

When an observation contains uncertain traits, the contributions from different observation patterns need to be accounted for. For example, the observation $\texttt{11*}$ can be matched by a signal with $\mathbf{s} = \{0\}$ emitting $\texttt{111}$ and by a signal with $\mathbf{s} = \{1\}$ emitting $\texttt{110}$. We then have

$$P_{obs}(\texttt{11*}) = P_\lambda(1) P(\mathbf{s} = \{0\}) P_{reach}(\texttt{111}) + P_\lambda(1) P(\mathbf{s} = \{1\}) P_{reach}(\texttt{110}) \tag{23}$$

$$= P_\lambda(1) \frac{1}{L} (P_{reach}(\texttt{111}) + P_{reach}(\texttt{110})), \tag{24}$$

where the second line follows because, in the case of single observations for each lineage, $P(\mathbf{s} = \{c\})$ is equal to $1/L$ for all c, as the probability of emission from any single state is uniform.

To illustrate the structure of probability calculations associated with emission probabilities, consider a system constrained to always follow the trajectory $\texttt{111} \to \texttt{011} \to \texttt{001} \to \texttt{000}$. The probabilities associated with several different observations, and the possible sets of states giving rise to each observation, are illustrated in Table 4.

Incomplete data

We have discussed the case of complete data observed in parallel and/or serial descents pattern, and incomplete data observed in parallel descent patterns. A natural question is how to deal with incomplete data in serial descent

bioRxiv preprint doi: https://doi.org/10.1101/037960; this version posted January 26, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Observation	**s**	$P_{emission}$	$\frac{P_{emission}}{P_\lambda(n)}$	Notes
001	$\{2\}$	$P_\lambda(1)\frac{1!}{1!}\frac{1}{L}$	1/4	Single state observation; one of four states in the trajectory matches this state.
00*	$\{2\}$ or $\{3\}$	$P_\lambda(1)\left(\frac{1!}{1!}\frac{1}{L}+\frac{1!}{1!}\frac{1}{L}\right)$	1/2	Two states, at different 'distances' (number of steps) from the origin, match this observation. The probability of reaching either of these states is accounted for.
***	$\{0\}$, $\{1\}$, $\{2\}$, or $\{3\}$	$P_\lambda(1)\left(\frac{1!}{1!}\frac{1}{L}+\frac{1!}{1!}\frac{1}{L}+\frac{1!}{1!}\frac{1}{L}+\frac{1!}{1!}\frac{1}{L}\right)$	1	Every possible signal matches this observation; hence, its observation probability is (intuitively) proportional to 1.
111 → 001	$\{0,2\}$	$P_\lambda(2)\frac{2!}{1!1!}\frac{1}{L^2}$	2/16	Here, two independent signals A and B are emitted from the trajectory. There are 16 possible pairs of emissions (A,B); two of these pairs give rise to the observation (specifically, A at 111 and B at 001, or A at 001 and B at 111).
111 → 011 and 011 → 001 (continuous)	$\{0,1,2\}$	$P_\lambda(3)\frac{3!}{1!1!1!}\frac{1}{L^3}$	6/64	Here, the two observations are assumed to be continuously linked, so the observation of 011 after the first transition immediately accounts for its presence at the start of the second; we thus only require three signals to be emitted.
111 → 011, 011 → 001, and 001 → 000 (continuous)	$\{0,1,2,3\}$	$P_\lambda(4)\frac{4!}{1!1!1!1!}\frac{1}{L^4}$	24/256	An extension of the previous coupled trajectory observation.
111 → 011 and 011 → 001 (not continuous)	$\{0,1,1,2\}$	$P_\lambda(4)\frac{3!}{1!2!1!}\frac{1}{L^3}$	3/64	In this example (in contrast to the picture described in the text, where continuity is assumed), the two observations are not assumed to be continuous, although they are assumed to come from the same lineage. We therefore require an extra signal to be emitted, characterising the system at the start of the second transition.

Table 4: **Probabilities of emission patterns for specific targets.** Here we demonstrate emission probabilities corresponding to different observations. We constrain evolution to follow the trajectory 111 → 011 → 001 → 000. This has the effect of removing transition probability factors from observation probabilities, as each transition has probability 1. The remaining observation probabilities thus depend only on the system emitting signals at appropriate points in the trajectory. For each observation, the pattern of signals **s** that give rise to that observation is shown, and the emission probability $P_{emission}$ from Eqn. 19 is given. We note that for more general transition networks, many more trajectories would be supported, and transition terms from each, not in general reducing to 1, would have to be included: accounting for these terms is the focus of the majority of the previous analysis.

patterns. We aim to address this more complicated question in further work; it is not required for the original application of this approach [46] (where evolution was completely parallel) or the application in this paper (where data is complete).

Several complications prevent the natural extension of the above analysis to lineages containing incomplete data. Firstly, ancestral properties inferred from incomplete contemporary measurements may themselves be incomplete. For example, if two descendents have properties *11 and 01*, the most reasonable inference of the properties of their ancestor is *11.

Secondly, when computing observation probabilities involving an initial state with incomplete data, we cannot be certain that that initial state has been reached by a previous transition in the lineage under consideration. Consider a lineage in which a step is known to have identified a state in the set $U(t)$. We are interested in the probability of observing a transition to b; recall that $P_{obs}(a\to b)=P(b|a;\pi)P_{reach}(a;\pi)$. If $a\in U(t)$, $P_{reach}(a;\pi)$ will be nonzero but in general not unity. We therefore need to consider $\sum_{a'\in U(t)}P_{obs}(a'\to b)=P(b|a';\pi)P_{reach}(a';\pi)$.

Thirdly, the patterns of signal emission are rather more complicated in the case of uncertain data. Previously, the structure of **s**, denoting the stages at which signals are emitted, was either fixed for each independent lineage (for example, Eqn. 22) or consisted of single elements (for example, Eqn. 23). For lineages involving several incomplete observations, we will in general have a large number of possible **s** structures, each involving several different elements. Furthermore, the time ordering of assumed signal emission times may vary according to the specific set of states being considered as responsible for those signals. Extending the above calculations to account for all possible emission pattern options in these cases will be the central focus of future extensions of this work.

For now, we note that the simplest possible incarnation of this general approach will be equally applicable to cases with incomplete data. That is, simulating evolutionary trajectories and random signal emission from each, then comparing the emission of signals to observed data, will yield an estimate of the likelihood associated with a set of observations. However, the simplifying steps applied in the above analysis, such as efficient path location using HyperTraPS, the neglection of emission probabilities as constant multiplicative factors, and assumptions of continuous time-ordered progression through lineage observations, may cease to hold in this more complicated case, and the straightforward simulation approach may thus be computationally demanding. Further work will identify simplifying approaches and efficient simulation protocols to use in this context.