nature portfolio

Peer Review File

Multiple distinct evolutionary mechanisms govern the dynamics of selfish mitochondrial genomes in Caenorhabditis elegans

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Authors: Gitschlag et al.
Manuscript Title: "Multiple distinct evolutionary mechanisms govern the dynamics of selfish mitochondrial genomes"
Tracking #: NCOMMS-24-06482-T

Summary:
This is a very nice analysis of multilevel selection on five mtDNA deletions in C. elegans using theoretical and empirical approaches. The authors use empirical data on intra- and inter-individual selection to model the frequency distribution of mtDNA deletions. The analysis suggests that the frequency distribution of mtDNA deletions in a population can depend primarily on frequency-dependent intra-individual selection, inter-individual selection and a combination of the two. The manuscript is clear and well-written for the most part (but see Minor Comments). The findings are noteworthy given a central gap in basic fundamental understanding of mitochondrial evolutionary dynamics and population biology despite their widespread used as molecular markers in molecular systematics.

Major Comments:

I have some questions about experimental procedures and how they might influence the results.

1. In the Methods, intra-individual selection appears to be estimated across a single generation by comparing the mutant mtDNA frequency in a parent to that of a pooled sample of three offspring. The stated purpose is that it reduces the effect of drift on the parent-offspring comparison. Could that influence the estimates of N, the mtDNA bottleneck since the strength of genetic drift is a function of N? Maybe I missed how the analysis corrects for the sample size in estimating N but it should be made clear.

2. In the selection experiment, the population sample is pooled and lyzed together. How is the frequency of the deletion measured in these samples if they are not performed on individuals? If this analysis tracks changes in the frequency of the deletion in pooled samples rather than changes in the frequency of individuals carrying the deletion, then the results reflect both intra- and inter-individual selection. What happens to mtDNA copy number under selection? Doesn't using pooled samples ignore any potential consequences in changes in mtDNA copy number? For instance, if individuals with a high intracellular frequency of an mtDNA deletion upregulate mtDNA copy number during the course of the experiment, could these individuals skew the results towards higher frequency of the mtDNA deletion in competition?

3. Could the authors provide more details in the Methods section with respect to crossing scheme used to place the mtDNA variants in the Bristol nuclear background? No information on the number

of generations of backcrossing is provided so it is hard to assess what proportion of the nuclear genome is expected to be wildtype. Also, how can the authors distinguish that progeny from a particular backcross are indeed from a hermaphrodite mating with a male, rather than a hermaphrodite selfing?

Some Comments (not major):
1. It is surprising that two of the deletions lack an apparent fitness cost in competition and also do not exhibit a detectable intra-individual advantage. Can these deletions be characterized as selfish?

2. It might be helpful to have a definition of N (the intra-individual bottleneck) in the legend of Figure 2. It is defined in the text but some readers might confuse it with individual population size out of habit.

3. Throughout the manuscript, the authors use the term "organismal selection." This can be confusing because it could also refer to "within-individual" or "intra-individual" selection. Care must be taken to delineate "intra-individual" versus "inter-individual" selection (or "within-individual and "between-individual" selection). It would be preferable if the authors use "inter-individual" or "between-individual" in lieu of "organismal."

The following are more minor points.
Minor Comments:
• Abstract, line 23 "Genetic drift" might be a better choice of phrase in lieu of "neutral drift."
• Introduction, line 65. Grammatical error. Correct "on one hand" to "on the one hand."
• Introduction, line 78. Replace "population genetic" with "population-genetic."
• Results, line 95. Replace "population genetic" with "population-genetic."
• Results, line 95. "Evolutionary dynamics" might be more appropriate than just "dynamics."
• Results, lines 98 and 128. Clarification. Do the authors mean "inter-organismal" level when they say "organismal?"
• Results, lines 129 and 130. Clarification. Do the authors mean "inter-organismal selection" when they say "intra-organismal selection?" It is not clear here if they are referring to (i) both selection and drift within an individual or (ii) selection between individuals and drift within an individual.
• Results, line 141. Replace "population genetic" with "population-genetic."
• Results, lines 153-154. The sentence "Thus, intra-organismal selection constitutively mutant frequency up into a range that elicits a strong organismal fitness cost (Fig. 2e)" is not making sense. Are the authors missing word(s)?
• Results, lines 155, 160, 172. Clarification. For clarity, it might help to replace "organismal" with "inter-organismal"
• Results, line 263. Replace "mid-size" with "mid-sized"
• Results, line 322. Replace "evade" with "evades"
• Results, line 335. Replace "closely related" with "closely-related."
• Methods, line 383. Replace "until being used" with "until use."
• Methods, line 445. How did the authors estimate a population size of 500 nematodes?
• Methods, line 453 and 454. Replace "population genetic" with "population-genetic."

• The References section is very sloppily done. The reference list needs to be heavily edited for formatting issues given the lack of consistency. Some article titles are listed with each word starting in uppercase, others not. In many instances, species names are not italicized.
• Figure 1 legend. The authors use the term “between-host.” For the sake of consistency, they ought to refer to this as the suggested “inter-individual.”
• Through the manuscript, please italicize “N” when it is used to denote population size.
• Figure 2. Please replace “neutral drift” with “random genetic drift” or “genetic drift.”

Reviewer #3:
Remarks to the Author:
The authors investigate the different dynamics by which mtDNA mutations evolve in nematode populations and, by linking detailed experiments with a population genetic model, connect these dynamics to different evolutionary mechanisms. They find that different mechanisms can best explain observed behaviour for different mtDNA mutations, demonstrating an interesting range of possibilities for mtDNA evolution.

I think this work is very interesting and a particularly nice demonstration of where modelling and detailed experiments can mutually reinforce to shed light on fundamental biology. This combination is a new approach for the particular question of selfish proliferation of mtDNA variants and the findings will have implications well beyond the particular study system. I have several comments about the implementation and one set of questions which to my eyes need resolving for full interpretation of the results. I'll lead with that (to me) most important point and follow up with some smaller-scale ones. NB -- this has turned out to be quite a long report, but please don't think the word count means there's a long list of issues. Most words are spent pinning down a quite technically involved set of questions about the statistics (which I do think are important, but may be down to my misunderstanding and/or may be easy to resolve).

For transparency, I am Iain Johnston and I am happy for this review to be treated as public domain. To my eyes my most important shortcoming as a reviewer here is a lack of experience with nematode lab work; I cannot comment on the husbandry, feeding competition, and any worm-specific physiological implications of these mutations.

-- Lead-order questions

I have several coupled questions about the modelling and bootstrapping.

First, for positioning -- we are doing parameteric bootstrap internally within each sample, right? So, for example, in Fig 3b, is it the case that each red line is actually a collection of 100 traces (one collection for each of the n=8 datasets), each of which comes from a computational re-simulation of that system under the maximum likelihood parameterisation from the original data?
If so -- I'm not sure of the connection between this and the task to "estimate our confidence in these parameter estimates" (l530). The within-sample parameteric bootstrap will give us a range of

parameter values for each sample, but the scientific results (as in Table 1) seem to be inferences about these values across different samples. How do the cross-sample values in the table come from the within-sample values from e.g. Fig 3b? Wouldn't it be reasonable to bootstrap-resample the dataset as a whole to get more traditional bootstrap confidence intervals on the parameters of interest?
Related -- do the bootstrap traces in the plots like 3def, 5a-l also correspond to sets of collections of within-sample resimulations, or is the approach here different?

Second and perhaps most importantly, I'm confused by the level of support from the bootstrap analysis for some model structures. For example, although it seems from the code that the gamma parameter determining the ascending/descending nature of w_intra with z is allowed to vary from -inf to +inf, we extremely rarely see ascending behaviour (a couple of traces in Fig. 5j) even when the data would seem quite ambiguous about the direction of the relationship. For example, in Fig. 5d, it is hard to see why increasing fitness with z wouldn't provide an equally good fit to the data -- and especially hard to see why no w_intra(z=0) values below ~2 are ever supported. Why not, for example, start at w_intra(0) = 0, ascend gently through the data around z = 0.8, and top out around 3?

I understand that the model is complex and multi-level, so that it may not be the case that the datapoints in these plots are the only observations contributing to the shape of the model fits. But if there are other observations that are constraining the fits towards these decreasing functional forms, it would be really nice to understand how this constraint works. To take that previous example, what is it about the mpt4 observations that mean lower w_intra(z=0) values are impossible?

My pessimistic concern is that there's something about the numerical fitting process that is artificially favouring parameterisations with the decreasing w_intra(z) trend, and therefore providing undue support to that region of parameter space. This is why I below (**) ask some questions about the initial condition dependence of the optimiser -- we always start with a particular gamma value (1), and it would be good to know that this is not biasing the results. If we start instead with a value that corresponds to the null hypothesis (0), or one with the opposite sign (-1), does the optimiser always identify the same solutions?

I appreciate that the model is validated with synthetic data -- but this doesn't in itself address the above, because if I am understanding correctly then all the synthetic data are generated to match this decreasing trend. Could the authors construct a synthetic set from a generator that has two alternatives -- (i) unchanging and (ii) increasing w_intra with z -- and show that the pipeline can equally well capture these behaviours?

Third, the bootstrap sets for some model elements (Fig. 3e, 5k) look quite multimodal -- ie two or more dense ensembles of traces separated by a sparse region. Why is this? Is it related to my first question about (re)sampling within samples as opposed to over the full dataset?

-- Smaller points

I'm not sure I agree with the positioning of l70-73. Certainly different study organisms will have different specific influences on mtDNA behaviour. But if the goal is to draw general conclusions about such behaviour, the solution cannot be to focus on a specific model. Rather the opposite -- a range of models is essential so that the specific features of each can be characterised and accounted for, and what remains can be classed as general. Confusingly, this is what the manuscript seems to suggest in l66-67, where the shortcomings of a single model focus are described.

Beginning l143, different mechanisms are outlined, illustrated in Fig. 2. The interplay between parameter requirements and resultant dynamics isn't super clear here. In the first one (l144-147) for example, we require organismal selection to be negligible below some frequency z*... but surely this is determined by the parameter choice for the selection function? Fig 2 would suggest this -- but the causality isn't clear. Is the story -- IF we are in this parameter regime THEN this mechanism holds?
Unless I am misunderstanding, the requirements on model parameters for each of these mechanisms to exist would be important to include here.

The mathematical methods section is, I think, pretty hard reading. It would help immeasurably to have a figure putting graphs and illustrations to the various expressions involved (and it'd help to label the equations!). e.g. how w_intra varies with gamma and epsilon, how w_org varies with alpha and beta. That could readily be an SI fig. How the various quantities involved (q, w_intra, etc) relate to observable quantities like mean heteroplasmies would also be very useful. This could readily be included via annotation (perhaps with some new content) in Fig 1c. l478-481 could be made much clearer by working with proportions rather than discrete numbers, i.e. setting z = i/N, 1-z = (N-i)/N. How w_intra influences the system isn't clear until we meet l478 -- perhaps the definition of q given inline in l481 could be promoted to where we first meet w_intra?

Referring to equations by their line numbers, in Eqn 505 we have a noise term e_intra, which is normally-distributed. But this would seem to disrespect the constraints on the variable to which it contributes, which is constrained on [0,1]. Do we risk getting nonsensical behaviours here, and would a constrained noise term be more appropriate?

I wasn't sure why a smoother -- particularly with a particular standard deviation -- was used on l519 (and again on l540). In some cases N=10 for example -- smoothing the discrete distribution with kernel width 0.1 would seem to give some probability of getting negative observations. Doesn't the discrete distribution already give a tractable likelihood for discrete (and appropriately constrained) draws?

-- ** Computational implementation -- more detailed notes

I had some questions about the computational implementation. I tried to run the Github code on the data provided in the review process, but as the manuscript system changed all the filenames of

the attached datasets I couldn't run the code and explore things myself. My first questions though are about the maximum likelihood process. We have a very nonlinear and quite highly parameterised model to fit so I wanted to explore the behaviour of the fitting process.

As far as I can see, the max_likelihood_values function is called twice, once from l528 and once from l331. On l331 it looks like the initial guess for the optimiser (the "itl" argument) is just the set of known parameter values (a, b, g, etc) that generated the synthetic data in the first place? And from l528 it looks like the same constant initial guess (1, 1, 1, etc) is used for every optimisation run?
I may well be misunderstanding -- apologies if so. But if this is the case, my two questions are:
1. If l331 is already passing the known true parameters as the initial guess to the optimiser, how can it be a fair test of the optimiser's ability to recover the true parameter values? Surely we should give it the same initial conditions as the "real-world" version (l528)?
2. How much do the optimisation results on l528 depend on the initial conditions?

Gitschlag *et al*., *reviewer comments* (italicized) with author responses (plain text)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Authors: Gitschlag et al.
Manuscript Title: "Multiple distinct evolutionary mechanisms govern the dynamics of selfish mitochondrial genomes"

Summary:
This is a very nice analysis of multilevel selection on five mtDNA deletions in C. elegans using theoretical and empirical approaches. The authors use empirical data on intra- and inter-individual selection to model the frequency distribution of mtDNA deletions. The analysis suggests that the frequency distribution of mtDNA deletions in a population can depend primarily on frequency-dependent intra-individual selection, inter-individual selection and a combination of the two. The manuscript is clear and well-written for the most part (but see Minor Comments). The findings are noteworthy given a central gap in basic fundamental understanding of mitochondrial evolutionary dynamics and population biology despite their widespread used as molecular markers in molecular systematics.

We thank the reviewer for these helpful comments and their support for the importance of our contribution here.

Major Comments:

I have some questions about experimental procedures and how they might influence the results.

1. In the Methods, intra-individual selection appears to be estimated across a single generation by comparing the mutant mtDNA frequency in a parent to that of a pooled sample of three offspring. The stated purpose is that it reduces the effect of drift on the parent-offspring comparison. Could that influence the estimates of N, the mtDNA bottleneck since the strength of genetic drift is a function of N? Maybe I missed how the analysis corrects for the sample size in estimating N but it should be made clear.

We have clarified this issue in the revised manuscript and added additional validation to show that our statistical procedure can accurately estimate *N*, by applying it to simulated data where the ground truth is known. These results are shared in the new Supplementary Figure 6.

The reviewer is absolutely correct that pooling the offspring reduces the variance in offspring allele frequency, which produces a better estimate of the form of intra-organismal selection but would produce a biased estimate of *N* if the variance in offspring mutant mtDNA frequency from this analysis were used naively. However, our inference procedure is not using the variance in mutant frequency from these parent-offspring comparisons to determine the strength of drift. Instead, the information about the strength of drift is coming from the shape of the stationary distribution of mtDNA frequencies. The deviations between the observed pooled offspring mtDNA frequencies and the expected frequencies due to intra-organismal selection are modeled phenomenologically by the error variance σ^2_{intra} (original manuscript lines 506-507) which in the model is not explicitly a function of *N*. Although

this approach leaves out the information from the parent-offspring comparisons in determining the strength of drift, our validation of our method on simulated data (Supplemental Figure 6) shows that our approach can accurately recover the ground-truth value, even though we are neglecting the information about drift from the parent-offspring comparisons. In the original manuscript, we included this validation for the other parameters but not for the validation for *N*. We apologize for this but have now included it in Supplementary Figure 6.

In order to be doubly sure that our method is not providing biased estimates of *N*, in the revision we also took a second approach where we instead explicitly modeled the variance in the parent-offspring experiment as reflecting the strength of genetic drift and performed the appropriate correction as suggested by the reviewer to account for pooling of progeny (equation 9 in the revised Methods). This produced estimates of *N* that are similar to our original method when applied to our empirical data (Supplementary Figure 7). We now include this additional validation (discussed in the Methods, lines 603-619 of the revised manuscript).

> *2. In the selection experiment, the population sample is pooled and lyzed together. How is the frequency of the deletion measured in these samples if they are not performed on individuals? If this analysis tracks changes in the frequency of the deletion in pooled samples rather than changes in the frequency of individuals carrying the deletion, then the results reflect both intra- and inter-individual selection. What happens to mtDNA copy number under selection? Doesn't using pooled samples ignore any potential consequences in changes in mtDNA copy number? For instance, if individuals with a high intracellular frequency of an mtDNA deletion upregulate mtDNA copy number during the course of the experiment, could these individuals skew the results towards higher frequency of the mtDNA deletion in competition?*

The reviewer raises a subtle yet important point related to the design of the competition experiment. We thank them for the opportunity to highlight this nuanced aspect in the updated manuscript.

The methodology of quantifying population-wide mutant frequency from pooled lysates, as opposed to sampling individuals during selection, was validated in a previous study from our group (Gitschlag et al. 2020 *eLife*), where we used both approaches on *uaDf5* and showed that they provided consistent estimates of selection. We have now clarified that we are relying on this already established methodology in the revised manuscript (lines 484-486 of the revised manuscript). Nevertheless, the point raised is important enough that it is worth expanding on here and in the manuscript.

While it is possible that the changes in mtDNA copy number in heteroplasmic animals can skew the heteroplasmy frequency in competed populations composed of heteroplasmic and homoplasmic wildtype animals, we account for this via the use of non-competed control populations. In essence, for each competition experiment that mixes heteroplasmic and wildtype animals, we also propagate non-competing control populations consisting of only heteroplasmic animals. We then divide the mutant mtDNA frequency in the competed populations at each generation by that of the non-competed controls, which corrects for changes in mutant mtDNA frequency that arise for reasons other than the presence of wildtype animals (in practice these changes in frequency are quite small). We have clarified the logic of these non-competed controls (lines 478-489 of the revised manuscript), which was not specifically addressed in the original version. Again, we thank the reviewers for the opportunity to clarify this important point.

> *3. Could the authors provide more details in the Methods section with respect to crossing scheme used to place the mtDNA variants in the Bristol nuclear background? No information on the number of generations of backcrossing is provided so it is hard to assess what proportion of*

the nuclear genome is expected to be wildtype. Also, how can the authors distinguish that progeny from a particular backcross are indeed from a hermaphrodite mating with a male, rather than a hermaphrodite selfing?

We acknowledge the importance of addressing the backcrossing method and thank the reviewer for emphasizing that this could be clarified in our manuscript. To ensure that our analysis was not confounded by variation in the nuclear genome, we completely exchanged the nuclear genome of each heteroplasmic strain with the nuclear genome of wildtype (Bristol strain) *C. elegans*, using a previously published unigametic inheritance method (Artiles, Fire, & Frokjaer-Jensen, 2019 *Dev Cell*). This method enables the complete replacement of the nuclear genome within two generations, by leveraging the activity of *gpr-1*, which encodes a G-protein regulator that regulates the forces exerted on the microtubules during mitosis. Over-expression of *gpr-1* increases the pulling forces on the pronuclei during prometaphase, resulting in the segregation of the paternal and maternal genomes into separate embryonic cell lineages. The germline of the hermaphrodite consequently inherits the nuclear genome of only one parent, allowing us to bypass the need for multiple generations of backcrossing. Thus, all heteroplasmic strains used in this study have identical nuclear backgrounds.

Briefly, each heteroplasmic strain was crossed to the *gpr-1* over-expression strain PD2220 following Mendelian genetics. Next, hermaphrodites of the stable *gpr-1* over-expression heteroplasmy strains were crossed to wildtype males. Non-Mendelian hermaphrodite progeny from these crosses, in which the paternal nuclear background is unigametically inherited in the germline cell lineage (determined by pharyngeal mosaic patterning), were individually propagated. Stock strains were established from the progeny of these animals, as they have a complete wildtype nuclear genomic background and retain the given heteroplasmy. This *gpr-1*-based strategy is in many ways superior to the repeated backcrossing strategy because it allows for a complete and clean swap of the nuclear genome. In contrast, with the repeated backcrossing strategy, there is a small possibility of retaining small fragments of the paternal nuclear genome. We have revised the Methods section (lines 380-386) to clarify this point.

Some Comments (not major):
1. It is surprising that two of the deletions lack an apparent fitness cost in competition and also do not exhibit a detectable intra-individual advantage. Can these deletions be characterized as selfish?

Although the empirical data (Figure 4) show no significant intra- or inter-organismal fitness effects for two of the five mutant mtDNA genotypes featured in this study, as the reviewer rightly notes, we argue in lines 331-333 of the revised manuscript that the maintenance of the heteroplasmic state is unlikely given a complete absence of any selfish advantage. This is because a genome with only neutral fitness effects would be expected to drift to fixation or extinction, rather than be stably maintained and co-transmitted alongside wildtype mtDNA. Consistent with this expectation, we further note in lines 333-335 that the maximum-likelihood results show a weak intra-organismal advantage when the mutant genomes are present in the low to mid frequency ranges. Finally, we note that although a deleterious impact on host fitness is a common feature, it is not a necessary characteristic of a selfish genetic element. In particular, the ability to ‘selfishly’ outcompete the wildtype genome, as well as the ability of the wildtype genome to regain an advantage as mutant frequency rises, do not necessarily require multilevel selection and may instead represent dynamics that occur entirely at a singular level, e.g. the intra-organismal level in the case of selfish mtDNA. This is consistent with prior reporting in other systems, in which the proliferation of ‘cheater’ entities is often limited by frequency-dependent selection. We cite reference 12 in lines 344-345 to highlight this point.

2. It might be helpful to have a definition of N (the intra-individual bottleneck) in the legend of Figure 2. It is defined in the text but some readers might confuse it with individual population size out of habit.

Thank you for the suggestion. We have added the definition to the legend.

3. Throughout the manuscript, the authors use the term "organismal selection." This can be confusing because it could also refer to "within-individual" or "intra-individual" selection. Care must be taken to delineate "intra-individual" versus "inter-individual" selection (or "within-individual and "between-individual" selection). It would be preferable if the authors use "inter-individual" or "between-individual" in lieu of "organismal."

Thank you for pointing out the confusion that can be caused due to the use of terminology. For better clarity, we have switched our terminology to "inter-organismal selection" and "organism-level fitness."

The following are more minor points.
Minor Comments:
• Abstract, line 23 "Genetic drift" might be a better choice of phrase in lieu of "neutral drift."
• Introduction, line 65. Grammatical error. Correct "on one hand" to "on the one hand."
• Introduction, line 78. Replace "population genetic" with "population-genetic."
• Results, line 95. Replace "population genetic" with "population-genetic."
• Results, line 95. "Evolutionary dynamics" might be more appropriate than just "dynamics."
• Results, lines 98 and 128. Clarification. Do the authors mean "inter-organismal" level when they say "organismal?"
• Results, lines 129 and 130. Clarification. Do the authors mean "inter-organismal selection" when they say "intra-organismal selection?" It is not clear here if they are referring to (i) both selection and drift within an individual or (ii) selection between individuals and drift within an individual.
• Results, line 141. Replace "population genetic" with "population-genetic."
• Results, lines 153-154. The sentence "Thus, intra-organismal selection constitutively mutant frequency up into a range that elicits a strong organismal fitness cost (Fig. 2e)" is not making sense. Are the authors missing word(s)?
• Results, lines 155, 160, 172. Clarification. For clarity, it might help to replace "organismal" with "inter-organismal"
• Results, line 263. Replace "mid-size" with "mid-sized"
• Results, line 322. Replace "evade" with "evades"
• Results, line 335. Replace "closely related" with "closely-related."
• Methods, line 383. Replace "until being used" with "until use."
• Methods, line 445. How did the authors estimate a population size of 500 nematodes?
• Methods, line 453 and 454. Replace "population genetic" with "population-genetic."
• The References section is very sloppily done. The reference list needs to be heavily edited for formatting issues given the lack of consistency. Some article titles are listed with each word starting in uppercase, others not. In many instances, species names are not italicized.
• Figure 1 legend. The authors use the term "between-host." For the sake of consistency, they ought to refer to this as the suggested "inter-individual."
• Through the manuscript, please italicize "N" when it is used to denote population size.
• Figure 2. Please replace "neutral drift" with "random genetic drift" or "genetic drift."

We thank the reviewer for the thorough reading and for these detailed suggestions. We have incorporated all the recommended changes.

Reviewer #2 (Remarks to the Author):

The authors investigate the different dynamics by which mtDNA mutations evolve in nematode populations and, by linking detailed experiments with a population genetic model, connect these dynamics to different evolutionary mechanisms. They find that different mechanisms can best explain observed behaviour for different mtDNA mutations, demonstrating an interesting range of possibilities for mtDNA evolution.

I think this work is very interesting and a particularly nice demonstration of where modelling and detailed experiments can mutually reinforce to shed light on fundamental biology. This combination is a new approach for the particular question of selfish proliferation of mtDNA variants and the findings will have implications well beyond the particular study system. I have several comments about the implementation and one set of questions which to my eyes need resolving for full interpretation of the results. I'll lead with that (to me) most important point and follow up with some smaller-scale ones. NB -- this has turned out to be quite a long report, but please don't think the word count means there's a long list of issues. Most words are spent pinning down a quite technically involved set of questions about the statistics (which I do think are important, but may be down to my misunderstanding and/or may be easy to resolve).

For transparency, I am Iain Johnston and I am happy for this review to be treated as public domain. To my eyes my most important shortcoming as a reviewer here is a lack of experience with nematode lab work; I cannot comment on the husbandry, feeding competition, and any worm-specific physiological implications of these mutations.

We appreciate the reviewer's transparency! We also thank them for these supportive and helpful comments, especially about the integration of modeling and laboratory experiments to shed light on fundamental biological questions, as showing the feasibility of this type of approach was one of the key motivations of our study.

-- Lead-order questions

I have several coupled questions about the modelling and bootstrapping.

First, for positioning -- we are doing parameteric bootstrap internally within each sample, right? So, for example, in Fig 3b, is it the case that each red line is actually a collection of 100 traces (one collection for each of the n=8 datasets), each of which comes from a computational re-simulation of that system under the maximum likelihood parameterisation from the original data?

We apologize that we have confused the reviewer about the overall structure of the bootstrapping scheme, and this has also led to some confusion about the interpretation of the bootstraps.

Figure 3 may be the root of this confusion because results of the parametric bootstrap are only shown in Fig 3e-f. The lines in 3a and 3b are least-squares fits to the individual experimental replicates. The idea is that 3a-c shows the "raw" data and then Fig 3e-f shows the inferred model and the bootstraps. We have clarified this in the Figure 3 caption of the revised manuscript. More generally, the bootstraps are not generated within each biological replicate but rather each bootstrap sample consists of newly sampled data for the entire set of experiments associated with any mitochondrial mutation.

If so -- I'm not sure of the connection between this and the task to "estimate our confidence in these parameter estimates" (l530). The within-sample parameteric bootstrap will give us a range of parameter values for each sample, but the scientific results (as in Table 1) seem to be inferences about these values across different samples. How do the cross-sample values in the table come from the within-sample values from e.g. Fig 3b? Wouldn't it be reasonable to bootstrap-resample the dataset as a whole to get more traditional bootstrap confidence intervals on the parameters of interest?
Related -- do the bootstrap traces in the plots like 3def, 5a-l also correspond to sets of collections of within-sample resimulations, or is the approach here different?

Again, we are absolutely re-sampling the data as a whole. Specifically, we infer the parameters of a generative model by maximum likelihood to produce a point estimate of the model parameters. We then sample new datasets, using the maximum-likelihood parameters, and rerun the inference on those resampled data to assess our confidence in the parameter estimates. These bootstrap samples show the distribution of model fits that would be produced under the assumption that our best fit model is accurate. Because maximum-likelihood estimation is consistent, in the large data limit our point estimates will converge to the true value and our bootstrap results will converge to the confidence intervals that would be calculated based on the Fisher Information under the standard asymptotic theory of maximum-likelihood inference. However, the parametric bootstrap provides a more accurate view of model performance outside the large data limit where the assumptions of asymptotic normality of parameter estimates may not hold (because we are simulating the inference under other plausible datasets rather than relying on normality assumptions that may not be true).

One question that may be in the back of the reviewer's mind is why we are doing a parametric bootstrap rather than some of the other types of bootstrap in common use (resampling data, resampling residuals, etc.). The main reason we are using the parametric bootstrap is because of the complexity of the experimental data, which includes the joint analysis of experiments with completely different designs and replication structures. Should we be resampling replicates, time points, reads within time points, etc.? The parametric bootstrap provides an elegant way to cut through this complexity.

Second and perhaps most importantly, I'm confused by the level of support from the bootstrap analysis for some model structures. For example, although it seems from the code that the gamma parameter determining the ascending/descending nature of w_intra with z is allowed to vary from -inf to +inf, we extremely rarely see ascending behaviour (a couple of traces in Fig. 5j) even when the data would seem quite ambiguous about the direction of the relationship. For example, in Fig. 5d, it is hard to see why increasing fitness with z wouldn't provide an equally good fit to the data -- and especially hard to see why no w_intra(z=0) values below ~2 are ever supported. Why not, for example, start at w_intra(0) = 0, ascend gently through the data around z = 0.8, and top out around 3?

The key insight is that the stationary distribution of mutant mtDNA frequencies also gives information about the nature of intra-organismal selection. Intuitively, the model can rule out patterns of selection that would produce stationary distributions whose shape is very different than the ones we observe.

In the case of models where intra-organismal fitness starts low (near or below neutrality) and increases with frequency, these produce stationary distributions that are qualitatively inconsistent with what we observe in the natural heteroplasmies (we now illustrate this point with the new Supplementary Figure 5r,u of the revised manuscript). Intuitively, a positive frequency-dependent intra-organismal fitness advantage means that heteroplasmic lineages with high mutant mtDNA frequency quickly show a substantial organismal fitness cost and are removed from the population by inter-organismal selection.

This results in stationary distributions that are peaked at low frequencies, contrary to what we observe in the empirical data (with perhaps the exception of *mptDf3*, where we do also see a negative value of gamma in one of the bootstraps).

> *I understand that the model is complex and multi-level, so that it may not be the case that the datapoints in these plots are the only observations contributing to the shape of the model fits. But if there are other observations that are constraining the fits towards these decreasing functional forms, it would be really nice to understand how this constraint works. To take that previous example, what is it about the mpt4 observations that mean lower $w_{intra}(z=0)$ values are impossible?*

In order to address the reviewer's point, in the revised manuscript we have added a supplemental figure to display the behavior of the model under a wider array of parameter values (Supplementary Figure 5), including examples similar to *mpt4* but with lower values of $w_{intra}(z=0)$ (Supplementary Figure 5p to 5u). We can see that these models show a completely different shape of stationary distribution than observed for *mpt4*.

> *My pessimistic concern is that there's something about the numerical fitting process that is artificially favouring parameterisations with the decreasing w_intra(z) trend, and therefore providing undue support to that region of parameter space. This is why I below (**) ask some questions about the initial condition dependence of the optimiser -- we always start with a particular gamma value (1), and it would be good to know that this is not biasing the results. If we start instead with a value that corresponds to the null hypothesis (0), or one with the opposite sign (-1), does the optimiser always identify the same solutions?*

Our optimizer is only conducting local optimization, and so we agree with the reviewer's concern that the initial condition could potentially be getting stuck in local optima. We have implemented the reviewer's suggestion, and for our empirical data we have rerun our maximum-likelihood inference procedure initialized at each qualitatively different solution from Supplementary Figure 5a, to ensure that we are not missing any important regions of parameter space, and used the parameter values from the highest likelihood achieved under any initialization. In addition, we have implemented some improvements to our optimizer (detailed below) in order to ensure that our optimization procedure is robust. None of these changes have changed the qualitative nature of our maximum-likelihood fits or our overall results and conclusions.

> *I appreciate that the model is validated with synthetic data -- but this doesn't in itself address the above, because if I am understanding correctly then all the synthetic data are generated to match this decreasing trend. Could the authors construct a synthetic set from a generator that has two alternatives -- (i) unchanging and (ii) increasing w_intra with z -- and show that the pipeline can equally well capture these behaviours?*

We have implemented this suggestion, and for each qualitatively different alternative scenario in our new Supplementary Figure 5a, we also now show that our inference pipeline can recover the ground-truth parameters (Supplementary Figure 5d to 5u).

> *Third, the bootstrap sets for some model elements (Fig. 3e, 5k) look quite multimodal -- ie two or more dense ensembles of traces separated by a sparse region. Why is this? Is it related to my first question about (re)sampling within samples as opposed to over the full dataset?*

The multimodality is indicating that a subset of bootstrap samples support a qualitatively different selective mechanism than the maximum-likelihood estimate. Being able to display this type of multimodality is one of the strengths of the parametric bootstrap over confidence intervals based on asymptotic normality. Taking the example of *uaDf5* (Fig. 3e) that the reviewer mentions, the multimodality indicates that the empirical data (Fig. 3b) are consistent with either a gradual or abrupt loss of organism fitness among animals with high heteroplasmic mutant mtDNA frequency, since both scenarios correspond to a similar mean organism fitness and produce a similar stationary distribution of mutant mtDNA frequencies. We have revised the manuscript to address this multimodality (lines 213-217 of the revised manuscript).

> *-- Smaller points*
>
> *I'm not sure I agree with the positioning of l70-73. Certainly different study organisms will have different specific influences on mtDNA behaviour. But if the goal is to draw general conclusions about such behaviour, the solution cannot be to focus on a specific model. Rather the opposite -- a range of models is essential so that the specific features of each can be characterised and accounted for, and what remains can be classed as general. Confusingly, this is what the manuscript seems to suggest in l66-67, where the shortcomings of a single model focus are described.*

We agree and have clarified this passage. Our main point here is simply that studying several heteroplasmies together in a single species and with a single modeling framework can provide insights that are complementary to those gained by the sustained focus on *uaDf5* or studying heteroplasmies that are in different species and hence not as directly comparable. We appreciate the importance of taxonomic diversity in mitochondrial biology that the reviewer mentions, and have revised the introductory section to better clarify the motivation for our study.

> *Beginning l143, different mechanisms are outlined, illustrated in Fig. 2. The interplay between parameter requirements and resultant dynamics isn't super clear here. In the first one (l144-147) for example, we require organismal selection to be negligible below some frequency z*... but surely this is determined by the parameter choice for the selection function? Fig 2 would suggest this -- but the causality isn't clear. Is the story -- IF we are in this parameter regime THEN this mechanism holds?*
> *Unless I am misunderstanding, the requirements on model parameters for each of these mechanisms to exist would be important to include here.*

We appreciate the need for clarity on this point and have revised the Figure 2 legend accordingly. The three mechanisms we describe are different a priori hypotheses for how a heteroplasmy could be maintained and do not depend on the specific choice of parameterization that we use for our inference. Essentially, a stable heteroplasmy can be maintained by (i) the intra-organismal dynamics having a stable fixed point $0<z^*<1$, (ii) inter-organismal selection opposing intra-organismal selection, or (iii) both mechanisms can be at play.

> *The mathematical methods section is, I think, pretty hard reading. It would help immeasurably to have a figure putting graphs and illustrations to the various expressions involved (and it'd help to label the equations!). e.g. how w_intra varies with gamma and epsilon, how w_org varies with alpha and beta. That could readily be an SI fig. How the various quantities involved (q, w_intra, etc) relate to observable quantities like mean heteroplasmies would also be very useful. This could readily be included via annotation (perhaps with some new content) in Fig 1c. l478-481 could be made much clearer by working with proportions rather than discrete numbers, i.e.*

setting z = i/N, 1-z = (N-i)/N. How w_intra influences the system isn't clear until we meet l478 -- perhaps the definition of q given inline in l481 could be promoted to where we first meet w_intra?

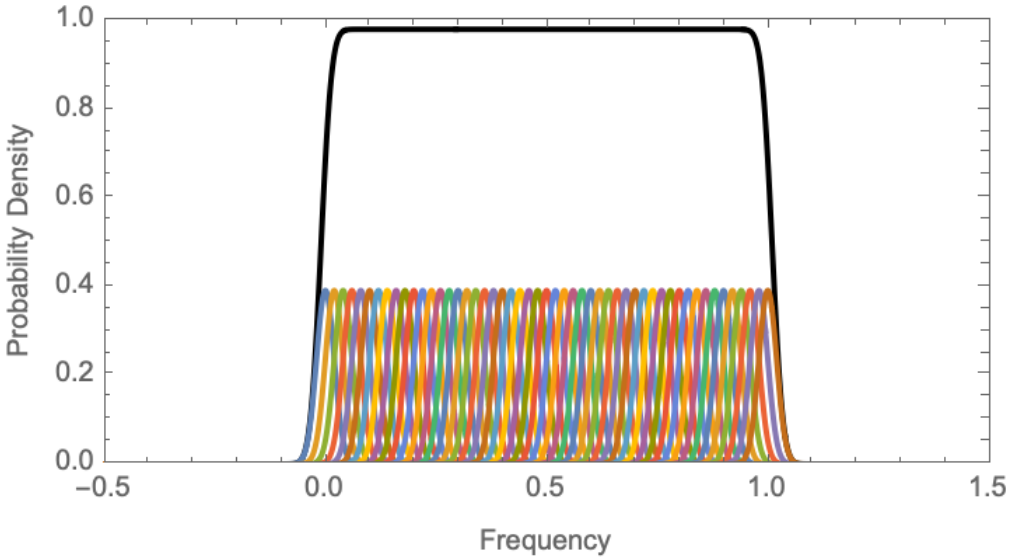
We have added a figure (the new Supplementary Figure 1) showing how the model parameters influence the form of the intra- and inter-organismal selection functions. We have also added numbers to the equations, as the reviewer suggested. The reviewer's suggestion for working with the continuous quantity z rather than i/N is certainly a good way to think about the model, but technically we need discrete indices i and j to indicate entries in the matrix M.

Referring to equations by their line numbers, in Eqn 505 we have a noise term e_intra, which is normally-distributed. But this would seem to disrespect the constraints on the variable to which it contributes, which is constrained on [0,1]. Do we risk getting nonsensical behaviours here, and would a constrained noise term be more appropriate?

In the revised manuscript we have implemented an alternative noise term which is constrained on [0,1] and show that it results in very similar parameter estimates for the empirical heteroplasmies (new Supplementary Figures 7-8). However, calculating this new noise term is too computationally intensive for our bootstrapping strategy (the likelihood for homoskedastic normal errors is extremely fast to calculate in terms of the sum of squared errors, whereas the more principled error requires a separate computation for each parent-offspring pair at each step of the optimizer), and so we retain the normally distributed errors for our main analysis. More broadly, while assuming normally distributed errors on frequencies means that our error model is necessarily misspecified, inference assuming normally distributed errors behaves similarly to fitting by least squares, which provides sensible behavior under a wide range of circumstances, and we show that it can accurately recover ground-truth parameters under simulation.

I wasn't sure why a smoother -- particularly with a particular standard deviation -- was used on l519 (and again on l540). In some cases N=10 for example -- smoothing the discrete distribution with kernel width 0.1 would seem to give some probability of getting negative observations. Doesn't the discrete distribution already give a tractable likelihood for discrete (and appropriately constrained) draws?

In the revised manuscript we have better explained the role and design of the smoother. Our experimental measurements of mutant mtDNA frequencies are continuous quantities whereas our model of genetic drift is based on a finite population size and hence produces a discrete distribution. The choice of bandwidth equal to the spacing of the discrete grid is a standard trick, in that narrower bandwidths produce smoothed densities that oscillate in the interior of the grid while wider bandwidths spill out further over the boundaries of the grid. For illustration, here is the performance of the smoother on the uniform distribution with N=50:



Uniform probability density function (black line) for *N*=50, obtained by summing *N* normal distributions (colored lines) spaced apart in increments of 1/*N*.

This is a reasonable approximation to the function that is 1 on [0,1] and zero elsewhere, and the bandwidth is optimal in the sense that narrower bandwidths result in a density that oscillates in the interior of the grid whereas wider bandwidths produce more spillover outside [0,1]. Although assigning probabilities to observations that cannot occur means that our model is necessarily misspecified, this misspecification is especially minor in our case because the stationary densities we are trying to estimate do not have much weight near either 0 or 1. This is because the stationary density is necessarily 0 at frequency 1 (all the mtDNA mutations we study eliminate at least one essential protein subunit of the electron transport chain) and stationary distributions with substantial mass near zero would result in a high rate of *de novo* loss of the heteroplasmy via genetic drift and hence would not be able to be maintained as heteroplasmic stocks. Finally, we have conducted extensive simulations showing that our procedure is sufficiently accurate to recover ground-truth values for a wide range of biological scenarios (Supplementary Figures 5-6).

*-- ** Computational implementation -- more detailed notes*

I had some questions about the computational implementation. I tried to run the Github code on the data provided in the review process, but as the manuscript system changed all the filenames of the attached datasets I couldn't run the code and explore things myself. My first questions though are about the maximum likelihood process. We have a very nonlinear and quite highly parameterised model to fit so I wanted to explore the behaviour of the fitting process.

As far as I can see, the max_likelihood_values function is called twice, once from l528 and once from l331. On l331 it looks like the initial guess for the optimiser (the "itl" argument) is just the set of known parameter values (a, b, g, etc) that generated the synthetic data in the first place? And from l528 it looks like the same constant initial guess (1, 1, 1, etc) is used for every optimisation run?
I may well be misunderstanding -- apologies if so. But if this is the case, my two questions are:
1. If l331 is already passing the known true parameters as the initial guess to the optimiser, how can it be a fair test of the optimiser's ability to recover the true parameter values? Surely we should give it the same initial conditions as the "real-world" version (l528)?

2. How much do the optimisation results on I528 depend on the initial conditions?

In the revised manuscript we consider the question of the initialization of the optimizer in more depth. For the point estimates of the empirical example, we initialize the optimizer at conditions corresponding to a wide variety of biological scenarios (the same intra-organismal fitness models in Supplementary Figure 5a). While many of these converge to the same values, for some initializations the optimizer does become stuck in local optima. In addition, we run the optimizer in a manner that updates the initializations. Specifically, for each step (corresponding to each value of *N* in the search space), the optimizer is initialized on both a neutral point (gamma=0, delta=2, epsilon=0), and also on the maximum-likelihood values from the previous step, and we use whichever optimization produced the higher likelihood. We then identify the overall maximum-likelihood parameters as those with the highest likelihood obtained from all approaches and initializations, which consistently corresponded to the values obtained by this latter recursive method, for all five genotypes.

For the bootstrapping procedure it is not feasible to consider a large number of initializations, and so we initialize the bootstrap replicates at the optimum from the empirical data. This increases the robustness of the calculation because the optimizer is already starting in a reasonable region of parameter space. While it is possible that the global optimum for any particular bootstrap replicate cannot be obtained from this initialization, the parametric bootstrap we are using is superior in this regard to the standard asymptotic maximum likelihood theory, since the parametric bootstrap relaxes the assumption of the asymptotic theory that the parameter estimates have a multivariate normal distribution around their true value.

Reviewer #3 (Remarks on code availability):

I have reviewed the code but could not get it to run on the data provided during the review process. This is not the authors' fault! The manuscript handling system renamed the datafiles to its own system -- serving no important purpose that I can see -- and the code relies on the original filenames for functionality. I hope the editor(s) can look into this as it is a substantial inconvenience for reviewing computational work.

I have therefore only been able to review the code by eye. If it is possible to provide the original datafiles (e.g. by email iain.johnston@uib.no<mailto:iain.johnston@uib.no>) -- under strict understanding of confidence of course -- I would be happy to look at it further.

(NB I have deliberately given my email address here as I have also signed my review)

We apologize that the code did not run, likely due to the renaming of the datafiles by the manuscript handling system as the reviewer points out. We have taken additional care in this resubmission to increase the likelihood that the file renaming by the submission website does not interfere with the ability of the reviewer to run the code.

Overall, we thank the reviewer for the thorough read of our manuscript and for putting in the effort to carefully think through the modeling. It has helped us improve the rigor and readability of the manuscript significantly.

Reviewers' Comments:

Reviewer #2

(Remarks to the Author)
Thanks very much for your explanations and for including this new material. This has cleared up all my questions about the methods and form of the fitting results and (I think) made some important details of the model rather clearer. I think the new Supp Figs 1 and 5 are particularly helpful for people who want to understand the range of behaviours the model *could* support and hence the subset of model behaviours most compatible with biological observation. I can't see that any other points need addressing and am happy to recommend this interesting and powerful study for publication. IJ

(Remarks to the Editor)

Reviewer #3

(Remarks to the Author)
1) The authors have made compelling arguments for their estimation of N in their analyses and have validated their method on simulated data. The procedure does not consider the variance in mutant frequency across generations in the first place, and therefore the sample size for parent-offspring comparison is not an issue. However, the authors also took a second approach where the parent-offspring variance was considered and obtained similar estimates of N.

2) The authors clarify the use of a pooled population and note that they normalise competed (heteroplasmic and homoplasmic WT) against non-competed (heteroplasmic only) populations in their analysis. Although this may not entirely correct for inaccurate heteroplasmy quantification in the samples (owing to changes in mtDNA copy in a manner dependent on heteroplasmy in individuals), it is practically realistic and I am satisfied with the explanation.

3) The authors provide sufficient detail in their Methods section on the use of unigametic inheritance method.

(Remarks to the Editor)
I believe the authors have done a good job in addressing the comments of Reviewer 1.

Gitschlag *et al.*, *reviewer comments* (italicized) with author responses (plain text)

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

*Thanks very much for your explanations and for including this new material. This has cleared up all my questions about the methods and form of the fitting results and (I think) made some important details of the model rather clearer. I think the new Supp Figs 1 and 5 are particularly helpful for people who want to understand the range of behaviours the model *could* support and hence the subset of model behaviours most compatible with biological observation. I can't see that any other points need addressing and am happy to recommend this interesting and powerful study for publication. IJ*

We thank the reviewer immensely for their valuable feedback, and for their support for our contribution with this study.

Reviewer #2 (Remarks on code availability):

I reviewed the code in some depth in the last round and the authors have addressed the points that arose. The authors have made the data available so the repo is now self-contained. It does not work out of the box because there is a path placeholder on l59 that doesn't exist on the user's machine: I suggest replacing this with
dir = '.'
so the code by default runs in the current working directory, while retaining the option for the user to change this.

I haven't in-depth stress-tested the code or reproduced every figure, but the code is running (after that edit) and producing what looks like sensible output. IJ

We appreciate the reviewer's concern to ensure that the code works out of the box. We have revised line 59 of the code to incorporate the reviewer's suggestion and have ensured that the code can be run from command line using the source data files available on our Github repository (linked in the Data Availability and Code Availability sections).

Reviewer #3 (Remarks to the Author):

1) The authors have made compelling arguments for their estimation of N in their analyses and have validated their method on simulated data. The procedure does not consider the variance in mutant frequency across generations in the first place, and therefore the sample size for parent-offspring comparison is not an issue. However, the authors also took a second approach where the parent-offspring variance was considered and obtained similar estimates of N.

2) The authors clarify the use of a pooled population and note that they normalise competed (heteroplasmic and homoplasmic WT) against non-competed (heteroplasmic only) populations in their analysis. Although this may not entirely correct for inaccurate heteroplasmy quantification in the samples (owing to changes in mtDNA copy in a manner dependent on heteroplasmy in individuals), it is practically realistic and I am satisfied with the explanation.

3) The authors provide sufficient detail in their Methods section on the use of unigametic inheritance method.

We greatly appreciate the favorable feedback from Reviewer 3, especially given their thorough read of our manuscript and the prior correspondence with other reviewers.

Overall, we have made diligent effort to ensure that all criticisms and questions raised by all reviewers have been rigorously and thoroughly addressed. We thank all reviewers once more for their helpful feedback and we are confident that incorporating their feedback has strengthened the findings of this study.