

Peer Review File

Manuscript Title: The ecological and genomic basis of explosive adaptive radiation

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

McGee, Seehausen and colleagues have, in conducting the most ambitious phylogenomic analysis of cichlid fish to date, enabled novel tests of the hypothesis that hybridization of divergent lineages has played a critical role in catalyzing the most explosive adaptive radiations—a hypothesis for which informative tests have been frustratingly elusive. Thus the work is novel, important, and of broad interest.

They first assemble a phylogeny of all described cichlids and reevaluate hypotheses supported by analyses previously focused mainly on sets of lake fish. Some variables important in earlier analyses fall away; aridity and the presence of large, visually hunting predatory fish prove the key (negative) factors. Multiple state speciation and extinction models suggest major rate differences among three sets of radiations, with the Lake Victoria Region radiating the fastest.

Seeking further insight into this variation, they conduct more detailed genomic analyses of ecologically divergent pairs of species from eight lake radiations, finding that more rapid speciation is associated with enrichment with often relatively ancient indel polymorphisms—which the authors interpret as evidence of the importance of early hybridization as a source of raw material for diversification. Different approaches to the analysis suggest the pattern is robust. Further, and remarkably, 214 indels in a related broad analysis were associated with diet, habitat or male nuptial coloration. One such pattern, concerning evolution of piscivory, is explored in more detail, and impressively. In my opinion it is the indel analyses and results that make the strongest case for the publication of this MS in an elite journal. I would be keen to see these findings explored further, perhaps including more about some of the other indel correlations and the genes possibly involved—even at the expense of the other analyses. In particular, is there a control analysis that can be conducted and presented for the results in Fig. 2b? What more can be done to confirm the hybrid origin of the indel polymorphisms in Fig. 2a?

The final analysis is nevertheless important (though it could easily go in a different MS). The authors analyze relationships among 100 Lake Victoria cichlid genomes using identity by descent blocks and networks. Their analyses suggest that habitat and diet diverged relatively early and contemporaneously, with male color patterns diverging repeatedly and more dynamically with speciation events, partially contradicting one popular conceptual model.

Detailed comments:

70-71: need some semi-colons

117: “sexually dimorphic too” seems lost

142: perhaps good to define “indels” somewhere early in the paper

147: this wording is tricky ... might “fixed between members of each species set” be more accurate? Or perhaps this is what they meant.

169: I wonder if choice of this frequency matters—i.e. would changing it affect the result?

178: what exactly does “across the outgroups” mean—in multiple outgroups? All? I am not sure and good to specify here rather than Methods.

257: To be fair though, Streelman and Danley also suggested differences in communication-related traits like color should be pronounced late in a radiation. Though my guess would be they were also involved in the early stages, but diverged and converged (and were sometimes lost) repeatedly over different speciation events, rather than just playing a role late.

269: better wording needed than “very unusually” ... maybe “extraordinarily”?

277: delete “so”

291-2: delete both or together

298: is regime the right word here?

347: “those of these” is awkward

Fig. 2 caption: Is it mentioned anywhere that the red symbols are the piscivores?

Fig. S3: I think there are columns here and each column is a species? Can this be clarified in the fig and/or the caption? I think “first” means most common?

Suppl 644: not really a sentence and hard to follow.

Suppl 761: independent

Suppl 778: but only

Suppl 788: significant p-values?

Suppl 1011: does “respectively” belong here? Maybe if substrate and diet are specified rather than ecology.

Table S6: could title row be added to 2nd part of table on 2nd page?

Referee #2 (Remarks to the Author):

This paper explores a very important question in evolutionary biology: what determines the rates of speciation. They found that varying rates in cichlids can be explained by the absence of/presence of predators, arid climate, and enrichment of old indel polymorphism. I agree that this is potentially interesting and important finding. However, I found several parts are unclear and cannot be completely convinced by the present analysis.

My major criticism is concerned with the analysis of genomic basis of rapid speciation (L137-161).

L139-140: How did you choose the two ecologically divergent species from each radiation? The choice of these species will substantially bias the subsequent analysis. I feel that the choice of only two species is not objective. I also wonder whether it is possible to conduct the same analysis using randomly sampled two species or two species with similar ecology. Is the extent of ecological divergence is proportional to the indel polymorphism?

L145: They assume that D_{xy}/π is proportional to time. Is this true? Which reference support this claim? Nei and Li (PNAS, 1979) suggest that d_A (d_{xy} - average π) can be proportional to time, but I do not know which paper claims that D_{xy}/π is proportional to time.

Minor comments;

L80: Why did you include the meristic characters for making the phylogenetic tree? Meristic characters

would be influenced by ecology. So, I suggest that they draw a phylogenetic tree solely based on DNA sequence data.

L90-91: Please show the DR of tropical freshwater and marine fishes.

L99: There is no Fig 1d.

L119-122: Which data support this claim (only haplochromine lineage underwent adaptive radiation)? Showing a phylogenetic tree of oreochromis only from Lake Victoria would be helpful. Alternatively, please cite any suitable reference.

L124-135: I think that the authors need to refer to Fig. S1 here. Otherwise, I cannot understand which data support this claim.

L222: How did you calculate the recombination rates?

Fig. 2c: Which species are shown here. What different colors indicate?

L473: How many bases of 26,058 bp are mitochondrial and nuclear?

L587: Before explaining bioinformatics methods, such as "variant calling pipeline", it would be better to explain how you obtained the short reads.

L730: Recombination map of what?

Author Rebuttals to Initial Comments:

Referee #1 (Remarks to the Author):

McGee, Seehausen and colleagues have, in conducting the most ambitious phylogenomic analysis of cichlid fish to date, enabled novel tests of the hypothesis that hybridization of divergent lineages has played a critical role in catalyzing the most explosive adaptive radiations—a hypothesis for which informative tests have been frustratingly elusive. Thus the work is novel, important, and of broad interest.

They first assemble a phylogeny of all described cichlids and reevaluate hypotheses supported by analyses previously focused mainly on sets of lake fish. Some variables important in earlier analyses fall away; aridity and the presence of large, visually hunting predatory fish prove the key (negative) factors. Multiple state speciation and extinction models suggest major rate differences among three sets of radiations, with the Lake Victoria Region radiating the fastest.

Seeking further insight into this variation, they conduct more detailed genomic analyses of ecologically divergent pairs of species from eight lake radiations, finding that more rapid speciation is associated with enrichment with often relatively ancient indel polymorphisms—which the authors interpret as evidence of the importance of early hybridization as a source of raw material for diversification. Different approaches to the analysis suggest the pattern is robust. Further, and remarkably, 214 indels in a related broad analysis were associated with diet, habitat or male nuptial coloration. One such pattern, concerning evolution of piscivory, is explored in more detail, and impressively. In my opinion it is the indel analyses and results

that make the strongest case for the publication of this MS in an elite journal.

This is an excellent concise summary of our results. We are pleased that the reviewer finds our results novel and strong.

I would be keen to see these findings (the indel analyses) explored further, perhaps including more about some of the other indel correlations and the genes possibly involved--even at the expense of the other analyses. In particular, is there a control analysis that can be conducted and presented for the results in Fig. 2b?

Control analyses for association studies are inherently problematic, as they involve comparisons among sets of extreme statistical outliers. However, we can test a more general model. We have added such an analysis to the Supplemental in section 2.4 showing that indels segregating in our outgroups exhibit stronger correlations with ecological traits than indels not present in any outgroup.

What more can be done to confirm the hybrid origin of the indel polymorphisms in Fig. 2a?

The relevant criterion here is the amount of indel variation relative to age, with explosive radiations associated with taxa that have diverged in far too many indels than they should have given their time since speciation. To this end, we have switched Figure 2a with its corresponding supplemental figure, emphasizing that the key feature here is unusually high levels of indel variation relative to the TMRCA of two lineages.

The final analysis is nevertheless important (though it could easily go in a different MS). The authors analyze relationships among 100 Lake Victoria cichlid genomes using identity by descent blocks and networks. Their analyses suggest that habitat and diet diverged relatively

early and contemporaneously, with male color patterns diverging repeatedly and more dynamically with speciation events, partially contradicting one popular conceptual model.

Detailed comments:

70-71: need some semi-colons

Fixed.

117: “sexually dimorphic too” seems lost

This paragraph has been moved to the Supplemental, and fixed.

142: perhaps good to define “indels” somewhere early in the paper

We have added a definition to the sentence.

147: this wording is tricky ... might “fixed between members of each species set” be more accurate? Or perhaps this is what they meant.

Yes, that is what we meant. We have changed to “fixed between the two species representing each radiation”

169: I wonder if choice of this frequency matters—i.e. would changing it affect the result?

We have relaxed the frequency to MAF 0.01 or greater, and results are qualitatively the same.

178: what exactly does “across the outgroups” mean—in multiple outgroups? All? I am not sure and good to specify here rather than Methods.

No; we have rephased to ‘present in one or more outgroup species’.

257: To be fair though, Streelman and Danley also suggested differences in communication-related traits like color should be pronounced late in a radiation. Though my guess would be they were also involved in the early stages, but diverged and converged (and were sometimes lost) repeatedly over different speciation events, rather than just playing a role late.

Yes we agree. Streelman and Danley proposed that these different classes of traits diversified in different stages of the radiation. Our analysis rather implies that all three classes of traits diversified early on, but that most major transitions between habitat and diet guilds happened quite early in the course of the radiation whereas transitions between the major male nuptial coloration types are very common even among the most recent speciation events.

269: better wording needed than “very unusually” ... maybe “extraordinarily”?

Changed.

277: delete “so”

Deleted.

291-2: delete both or together

Deleted.

298: is regime the right word here?

Yes, but we have added ‘rate regime’ as this is the more common phrasing.

347: “those of these” is awkward

Deleted.

Fig. 2 caption: Is it mentioned anywhere that the red symbols are the piscivores?

We have added a legend indicating this.

Fig. S3: I think there are columns here and each column is a species? Can this be clarified in the fig and/or the caption? I think “first” means most common?

We have changed the orientation and added clarification text to the caption.

Suppl 644: not really a sentence and hard to follow.

Section removed after change to Fig2a, as per Reviewer 2.

Suppl 761: independent

Fixed.

Suppl 778: but only

Fixed.

Suppl 788: significant p-values?

Clarified.

Suppl 1011: does “respectively” belong here? Maybe if substrate and diet are specified rather than ecology.

Clarified.

Table S6: could title row be added to 2nd part of table on 2nd page?

Done.

Referee #2 (Remarks to the Author):

This paper explores a very important question in evolutionary biology: what determines the rates of speciation. They found that varying rates in cichlids can be explained by the absence of/presence of predators, arid climate, and enrichment of old indel polymorphism. I agree that this is potentially interesting and important finding. However, I found several parts are unclear and cannot be completely convinced by the present analysis.

My major criticism is concerned with the analysis of genomic basis of rapid speciation (L137-161).

L139-140: How did you choose the two ecologically divergent species from each radiation? The choice of these species will substantially bias the subsequent analysis. I feel that the choice of only two species is not objective. I also wonder whether it is possible to conduct the same analysis using randomly sampled two species or two species with similar ecology. Is the extent of ecological divergence is proportional to the indel polymorphism?

The reviewer presents two issues. One, that our choice of species could be a source of bias. Two, that we only use two species per lake radiation, when we could have used more. We agree that both issues are important and worthy of additional explanation.

Regarding the choice of species, we have in each case selected species with maximal ecological and ecomorphological divergence within the radiation in question, first maximizing trophic level divergence (eg. comparing a fish-eating species to an herbivorous species, if the radiation possesses both ecotypes), and next by maximizing functional divergence (species that consume prey by suction feeding vs by biting) followed by habitat divergence and divergence in overall body shape. We believe that our choice of species can be clearly justified on these grounds, and have added a section discussing our choices in the Supplementary Information in section 2.2.

Regarding the choice of only two species per lake, we did this precisely to avoid biases associated with comparing clades with many species and clades with fewer. For example, our smallest radiation, the *Coptodon* species flock of Lake Ejagham, contains less than five species. Several of the endemics possess generalist diets typical of their riverine ancestors, but two species display a greater degree of ecological divergence. The radiations of Malawi and Victoria contain hundreds of potential species to choose from, but incorporating many of these species into the analyses would likely increase discovery of new indel variation to even greater levels than we observe in our current analyses, and correcting for the increased number of comparisons would not be straightforward

in the presence of phylogenetic structure within each radiation. We believe the safest and most consistent way to analyze across cichlid radiations of variable species richness is to follow our current strategy of sampling two highly ecologically divergent taxa per radiation. We have added a discussion of this issue to the Supplementary Information in the section “Genomic basis of speciation rate variation”, and thank the reviewer for highlighting the issue.

L145: They assume that D_{xy}/π is proportional to time. Is this true? Which reference support this claim? Nei and Li (PNAS, 1979) suggest that d_A (d_{xy} - average π) can be proportional to time, but I do not know which paper claims that D_{xy}/π is proportional to time.

The reviewer is correct that D_{xy}/π (absolute sequence divergence relative to polymorphism) as a means of measuring relative divergence time has less of a history in the population genetics literature than d_A , though we note that d_A did not become commonly utilized for that purpose either, despite Nei’s suggestion of its proportionality to time. We have therefore replaced Fig2a with its corresponding figure in the Supplemental, which compares speciation rate with the ratio of indels relative to TMRCA for the two focal species representing each radiation. This shows at least as strong a pattern and allows for simplification of the main text. We thank the reviewer for the suggestion.

Minor comments;

L80: Why did you include the meristic characters for making the phylogenetic tree? Meristic characters would be influenced by ecology. So, I suggest that they draw a phylogenetic tree solely based on DNA sequence data.

We agree that meristic characters can be influenced by ecology, for example, in traits such as gill raker counts. However, in this case, we were careful to only utilize meristic traits likely (and known from cichlid systematic taxonomy) to be more neutral. As an example, spine and fin ray counts are often identical or very similar between ecologically divergent members of young cichlid radiations in Victoria and Malawi, whereas ecologically equivalent species belonging to genetically divergent cichlid clades often are very different in meristic traits.

We have now added an additional analysis to the supplemental results, establishing extremely strong phylogenetic signal in each of our meristic traits, which can be found in section 1.4 of the Supplemental. Briefly, we produce a tree utilizing only our DNA sequence data (1000 out of 1712 species) but with all pipeline steps otherwise identical, then test the phylogenetic signal of each meristic trait. We find lambda values over 0.9 for all traits, indicating extremely high phylogenetic signal and suggesting that the traits can indeed be utilized for phylogenetic reconstruction. Using the meristic traits importantly allows us to place the 712 described species for which DNA data is not available.

L90-91: Please show the DR of tropical freshwater and marine fishes.

We have opted to remove this sentence in order to reduce word count, as a comparison of cichlids to other fish clades is not a main focus of the paper.

L99: There is no Fig 1d.

Fixed.

L119-122: Which data support this claim (only haplochromine lineage underwent adaptive radiation)? Showing a phylogenetic tree of oreochromis only from Lake Victoria would be helpful. Alternatively, please cite any suitable reference.

We have moved this the Supplemental in section 1.8 for space reasons, and added a citation for the claim (Muschick et al 2018).

L124-135: I think that the authors need to refer to Fig. S1 here. Otherwise, I cannot understand which data support this claim.

We have added citations to the Supplemental for both our HiSSE and MiSSE results.

L222: How did you calculate the recombination rates?

Recombination rates are from a previously published linkage map cited in the Supplemental (Feulner et al 2018). We have added a clarifying statement in the Supplemental: "...using the recombination map from the most recent Pundamilia nyererei assembly to convert from bp to cM."

Fig. 2c: Which species are shown here. What different colors indicate?

We have added additional legends explaining the piscivore/insectivore ecologies, and added species names to the caption.

L473: How many bases of 26,058 bp are mitochondrial and nuclear?

We have added this information to the Supplemental: "(sixteen mitochondrial loci (15,005 bp) and thirteen nuclear protein coding genes (11,589 bp))".

L587: Before explaining bioinformatics methods, such as "variant calling pipeline", it would be better to explain how you obtained the short reads.

We clarify this at the beginning of section 2.1 in the Supplemental.

L730: Recombination map of what?

We have added this information: "the recombination map from the most recent Pundamilia nyererei assembly to convert from bp to cM."