nature research

Peer Review Information

Journal: Nature Ecology & Evolution
Manuscript Title: Balancing selection maintains hyper-divergent haplotypes in C. elegans
Corresponding author name(s): Erik C. Andersen

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

9th November 2020

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Erik,

Thank you for your patience while we waited for comments on your manuscript entitled "Balancing selection maintains ancient genetic diversity in C. elegans". The manuscript has now been seen by three reviewers, whose comments are attached. The reviewers have raised a number of concerns which will need to be addressed before we can offer publication in Nature Ecology & Evolution. We will therefore need to see your responses to the criticisms raised and to some editorial concerns, along with a revised manuscript, before we can reach a final decision regarding publication.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file in Microsoft Word format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

* Include a “Response to reviewers” document detailing, point-by-point, how you addressed each reviewer comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the reviewers along with the revised manuscript.

* If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions at http://www.nature.com/natecolevol/info/final-submission. Refer also to any guidelines provided in this letter.

* Include a revised version of any required reporting checklist. It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review. A revised checklist is essential for re-review of the paper.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Ecology & Evolution or published elsewhere.

Nature Ecology & Evolution is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on ‘Modify my Springer Nature account’. For more information please visit please visit www.springernature.com/orcid.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

[REDACTED]

Reviewer expertise:

Reviewer #1: evolutionary genetics and comparative genomics

Reviewer #2: evolutionary genetics and comparative genomics

Reviewer #3: evolutionary genetics and comparative genomics, Caenorhabditis

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The manuscript from Lee et al. describes a population genomics re-analysis using whole genome sequencing data from hundreds of C. elegans samples and new assemblies of additional whole genomes from several collections. The authors argue for a center of diversity for C. elegans in the pacific islands, but that most of the C. elegans genome has very low levels of genetic diversity. However, as has been observed in other species with small effective population sizes, the authors find some regions of the genome with much higher density of polymorphism. The authors propose that these regions harbor ancient diversity maintained by stable balancing selection, and find that they are enriched for genes that play a role in environmental response.

The finding that highly diverse regions of the genome play a major role in shaping genetic diversity in C. elegans in is an exciting continuation of the trend identified in several recent studies of selfing plants. Furthermore, the de novo assembly of many such regions is a major step forward in understanding such regions. Linking this diversity to actual phenotypic variation is also an important step forward. However, I have several concerns about the statements made in the manuscript and some of the analyses. Most importantly it is unclear to me whether the authors have found sufficient evidence to justify the conclusion that balancing selection has driven retention of alleles, and I don't think that the authors have shown that these alleles are ancient. I think that, even without these two assertions, the manuscript is exciting, and if at least evidence for balancing selection could be strengthened then the manuscript would be considerably improved. Please see my more detailed comments below.

The manuscript does not have any introductory material, and instead starts immediately with results. I recognize that this is a short format however, at minimum, the authors should provide a paragraph describing previous evidence of ancient diversity being maintained by balancing selection (both the classical cases and those recently identified by genomic scans, some of these studies are referenced later but should probably be in a short introduction as well), and another paragraph explaining the current understanding of global C. elegans diversity, distribution, and species-wide. The addition of an introduction might help with some of the problems highlighted below.

The claim of origin of C. elegans in the Pacific seems to be based on the same dataset as published in Crombie et al. according to the listed project number (PRJNA549503). I certainly have no objection to a reanalysis of a previous dataset, however in my opinion it is not clear enough that the previous work, based on complementary methodology, came to many of the same conclusions. I think it is important to state this upfront, and to make it clear which parts of the results are new and which are confirmatory.

The argument for a Pacific origin is a little uncertain in my opinion. First, it would good to define the geographic region meant by "Pacific" clearly. For example, the authors might mean the western coasts of the Americas, the Pacific Islands, and/or eastern and southeastern Asia. Second, Hawaii is an international tourist destination, has extensive importation of crop species, and has a population consisting of with many immigrant peoples. High variation may simply reflect immigration of C. elegans populations from places not sampled by the authors. It is my opinion that this point should be

mentioned along with the authors very limited sampling in Asia.

If the authors decide to keep the early focus on population structure, it would be informative to understand the global distribution of isotypes. Were they always found in the same location?

Given the likely extended LD in C. elegans, the authors should explore downsampling their variant datasets for PCA to limit overestimation population divergence based on dependent, linked polymorphism. In addition, the authors should provide variance explained on each axis.

The authors assert that the several major sweeps in the C. elegans genome that they previously identified may have occurred during the out of the Pacific event. I think that Line 102-103 should read Extended Data Fig 2. It would be valuable to see all of the chromosomes rather than just these selected chromosomes for comparison sake. Also the beeswarm plot, presented by chromosome would be a helpful addition.

The authors move on to the major focus of the manuscript, the role of balancing selection in determining genomic diversity in genome of C. elegans. They begin by observing regions with exceptionally high values of diversity using standard estimators. They refer to these as high divergence, but I think using statistics like theta and pi it is better to use the term diversity. It should be noted that the authors use very small window sizes in these scans (1kb). What is the length of LD in C. elegans and does this make sense? Did the authors screen windows with only small numbers of accessible bases? Supplementary figure 4 is listed of evidence of poor alignment, but perhaps it would be better to look at alignment statistics in these windows directly rather than showing two examples.

The authors next use a set of empirical thresholds to define hyper divergent regions with whole genome assemblies and their illumina data. They describe a procedure by which they chose their empirical thresholds, but do not provide any data supporting their use of their thresholds or any of the results from their optimization. The authors should provide a clear justification for the threshold using the datasets from different thresholds and by showing where these particular thresholds fall in the overall distributions of diversity statistics/alignment statistics. How did the authors figure out what is a true divergent region to use to assess false positives and negatives?

The authors identify 20% of the genome that is hyper-divergent from the reference genome in at least one accession. This use of reference polarization makes some of the subsequent analyses difficult to interpret. For example, I am not sure why it makes sense to calculate the fraction of the genome that is hyperdivergent from the reference in each line. Does this not simply reflect the relationship to the reference line? Perhaps the authors intend to show which regions harbor the highest number of rare divergent alleles?

The divergent alleles identified by the authors are very interesting. There are a number of possible explanations for such alleles including random chance, introgression and balancing selection. The authors immediately suggest balancing selection as their preferred hypothesis, but fail to test any other hypothesis. The evidence they present is a increased Tajima's D at these loci. While this result is consistent with their hypothesis, Tajima's D is sensitive to many demographic factors and should be supported with additional tests. Also, large sections of the genome are argued by the authors to have been subject to directional selection, so wouldn't neutrally evolving loci also show elevated genomic Tajima's D if preselected to carry some genomic variation (as is the case here)? Also, how do the authors think that increased LD fits with their hypothesis of balancing selection? Long LD can occur at

balanced loci under some models, but certainly not all, so perhaps the authors should provide some discussion of this.

The authors do not attempt to understand the diversity of these regions in the context of the species demography, nor to formally model if they might occur by chance given a highly diverse outcrossing ancestor.

In the gwas, why was the overlap calculated post hoc rather than using the genotype of each line in the divergent region in the gwas? Obviously, kinship/relatedness should still be accounted for in such an analysis.

The authors do not show that these alleles are ancient (which sets up the claim of long term balancing selection). The argument for these alleles being ancient is that they are highly diverged, and that these divergences are similar to those observed between species in a related lineage. I think this result is suggestive (assuming that the error rates are the same in these regions as in other regions), but this can not be seen as a formal dating of the alleles as implied by the term ancient. There approach does not exclude variation in recombination rates, mutation rates, or introgression as contributing factors. If the authors identify some of these alleles in related species, that would provide evidence for ancient balancing selection. Otherwise, I suggest that the authors instead argue for balancing selection, and place the high divergences in context of previous studies that could compare alleles across species.

Reviewer #2 (Remarks to the Author):

This is an extremely well written MS that will be of great general value to the broader scientific community. This MS describes tracks of old polymorphism within C. elegans sampled world-wide, an argues that these tracks of high genetic diversity have been maintained by balancing selection. I have a few suggestions to the authors that I believe will substantially improve the plausibility and interpretability of their results.

1. Throughout the manuscript the authors use the phrase "haplotypes" to both describe genetically unique isolates and to describe patterns of genetic diversity at specific loci. As someone who doesn't think about C. elegans all the time, I had to remember that these wild isolates are basically inbred due to selfing. It took me a while to recognize that I was conflating my use of the word haplotype (which I typically think of as being restricted to a specific region of the genome) and the way the authors were using it (at least in some places). I think that the first confusing instance of the phrase haplotype is on line 78 ("We identified 328 distinct genome-wide haplotypes (henceforth, referred to as isotypes)"). The next instance of haplotype is the genetically localized on on line 103. It is a minor detail but I think that making the language more clear would make the paper more accessible to the outside reader.

2. I am skeptical of the GWAS and the enrichment test that the authors did. I recognize that the GWAS approach that the authors performed follows an established pipeline that is designed to greatly reduce the multiple testing burden. While I think that the approach is reasonable, I have a set of suggestions: First, the enrichment test you performed and reported as the result of a hypergeometric test is problematic for a number of reasons: (1) it is unclear if the test is driven by one or more than

one of the GWAS results (there were two or three phenotypes); (2) the hypergeometric test might not be the most appropriate given that the assumption of independence is violated with correlated signal across the genome - you could use any number of the overlap type tests that have been designed for genomic regions; (3) no actual results of the GWAS are given (like regions, p-values, etc, etc) and all we know is that the results beat a non-conservative implementation of Bonferroni. To solve issues #1 and #2, I think that the authors need to implement a permutation test where they permute the phenotype data, repeat the GWAS, and the enrichment test. If their actual data beat the permutation in terms of enrichment, then I will believe the result. If the GWAS do not beat the permutation then I will think that the result is artifactual and driven by an inappropriate use of multiple testing correction, enrichment test, or both. Also, the authors should report the results of their GWAS rather than point the reader to an technical supplemental table that I couldn't really understand (there are more than three columns/phenotypes, why?)

3. I really want to know what all those extra genes in the hyper-divergent are. Are they paralogs of other neighboring genes? For instance, are we observing the "accordion" genome effect where organisms duplicate and diversify genes to accommodate different functions? Or, are they molecularly unique? Any indication would be super informative and is conspicuously absent from a discussion.

Other than that, the paper was a pleasure to read.

Reviewer #3 (Remarks to the Author):

In this report, Lee et al use a combination of resequencing, GWAS on fitness-related traits to growth on different natural bacterial food sources, and analysis of previously published transcriptomics to characterize balancing selection in wild strains of C. elegans. Perhaps due to its mating system, study of C. elegans has led to the publication of a number of impactful reports characterizing the effect of various evolutionary forces on segregating genetic variation (e.g. Rockman et al, 2010 Science, Andersen et al, 2012 Nature Genetics). Here, the authors use short and long-read resequencing to study the effects of balancing selection in wild populations of C. elegans.

This work builds on a number of previous reports that has implicated the importance of balancing selection in maintaining genetic variation in C. elegans. In three previous reports, specific loci that affect responses to natural products – avermectins and pheromones – were identified (Ghosh et al 2012 Science, Greene et al 2016 Nature, Lee et al, 2019 Nature E&E). Each of these loci showed signatures of balancing selection using Tajima's D and two of these loci showed a remarkable amount of nearby genetic variation consistent with the action of long-term balancing selection. Two additional reports have identified toxin-antitoxin elements that show signatures of balancing selection, although it is unknown if these toxin-antitoxin elements are under balancing selection or linked to loci that are (Seidel, 2008 Science, Ben-David et al 2017 Science). Finally, resequencing of a C. elegans wild strain identified many remarkably divergent regions throughout the genome (Thompson et al 2015, Genetics). The high amount of variation is consistent with balancing selection retaining ancient genetic variation conferring functional changes to nearby genes, similar to human haplotypes in the MHC locus.

Lee et al build upon these results using resequencing of wild strains of C. elegans using a combination of short-read resequencing (609 strains) and long-read sequencing (14 strains). From this data, they

identify 366 new hyper divergent regions. Are these hyper divergent regions under balancing selection? They address this in three ways. 1) They use population genetics to show these regions have higher Tajiama's D than the rest of the genome 2) They show these regions are enriched for genes that are involved in environmental sensation, consistent with environmental heterogeneity. 3) They use GWAS and transcriptomics to characterize the response to natural bacteria and pathogens and show they overlap with these regions.

Overall, this paper should be of interest to evolutionary biologists – broadly to those interested in the role of balancing selection in maintaining genetic diversity in a population and the effect of mating system on evolutionary trajectories. It should also be of interest to the study of C. elegans evolutionary biology.

Some specific comments on the manuscript:

1. I'm not sure if this is intentional but there is no Introduction section. This would probably be useful to more formally introduce their working model for C. elegans evolutionary trajectory that frames the work. 1. Male/Female species in the Hawaiian islands with genetic diversity on chromosomal arms shaped by background selection. 2. Evolution to androdioecy followed by loss of genetic diversity except for regions under balancing selection. 3) Recent spread of C. elegans across the globe accompanied by selective sweeps across multiple chromosomes. It might be useful to include a figure panel that summarizes this model. However, this is an editorial decision for the authors to make.
2. Figure 1 – It is difficult to distinguish the blue colors for Hawaiian vs Atlantic strains.
3. Figure 2c – Are the authors underestimating the amount of genetic diversity that is contained within these regions as short read sequences are unable to align fully in these regions? Further, why are there two clusters of strains? Is there a reason not to follow their previous color coding by geographical origin?
4. Figure 2d – Why is there more genetic diversity in Atlantic populations? Is this consistent with the out of the Pacific origin for C. elegans? The authors indicate that these strains show less of the swept haplotypes. Would that be enough to explain the increased diversity?
5. I found it very confusing how the hyper divergent regions were defined. I understand parameters were chosen to maximizing overlap between long and short read technologies but it was unclear where the original definition is. Can the authors justify more what defines a hyper divergent region?
6. I think it is important to summarize more of the results from the long-read sequencing. How similar are the divergent regions produced by the long-read sequencing vs. small-read sequencing. How many large deletions/insertions are missed? How do they affect gene function? How many of the missing/added genes have homologs in nematodes and other species. Do any have homology to known toxin-antitoxin genes?

**********END**********

Author Rebuttal to Initial comments

Reviewer 1

The manuscript from Lee et al. describes a population genomics re-analysis using whole genome sequencing data from hundreds of C. elegans samples and new assemblies of additional whole genomes from several collections. The authors argue for a center of diversity for C. elegans in the pacific islands, but that most of the C. elegans genome has very low levels of genetic diversity. However, as has been observed in other species with small effective population sizes, the authors find some regions of the genome with much higher density of polymorphism. The authors propose that these regions harbor ancient diversity maintained by stable balancing selection, and find that they are enriched for genes that play a role in environmental response.

The finding that highly diverse regions of the genome play a major role in shaping genetic diversity in C. elegans in is an exciting continuation of the trend identified in several recent studies of selfing plants. Furthermore, the de novo assembly of many such regions is a major step forward in understanding such regions. Linking this diversity to actual phenotypic variation is also an important step forward. However, I have several concerns about the statements made in the manuscript and some of the analyses. Most importantly it is unclear to me whether the authors have found sufficient evidence to justify the conclusion that balancing selection has driven retention of alleles, and I don't think that the authors have shown that these alleles are ancient. I think that, even without these two assertions, the manuscript is exciting, and if at least evidence for balancing selection could be strengthened then the manuscript would be considerably improved. Please see my more detailed comments below.

The manuscript does not have any introductory material, and instead starts immediately with results. I recognize that this is a short format however, at minimum, the authors should provide a paragraph describing previous evidence of ancient diversity being maintained by balancing selection (both the classical cases and those recently identified by genomic scans, some of these studies are referenced later but should probably be in a short introduction as well), and another paragraph explaining the current understanding of global C. elegans diversity, distribution, and species-wide. The addition of an introduction might help with some of the problems highlighted below.

We thank the reviewer for this concern. We have added an introduction section to our manuscript which includes a paragraph describing previous evidence of ancient diversity being maintained by balancing selection and a paragraph about the current understanding of *C. elegans* genetic diversity.

The claim of origin of C. elegans in the Pacific seems to be based on the same dataset as published in Crombie et al. according to the listed project number (PRJNA549503). I certainly have no objection to a reanalysis of a previous dataset, however in my opinion it is not clear enough that the previous work, based on complementary methodology, came to many of the same conclusions. I think it is important to state this upfront, and to make it clear which parts of the results are new and which are confirmatory.

The dataset used in our manuscript is not the same as Crombie *et al.* 2019, as it contains an additional 103 strains, which were isolated across the world (Australia, Brazil, Czechia, France, Germany, Israel, Kenya, Netherlands, New Zealand, Peru, São Tomé and Príncipe, Spain, and United States). The data for these

additional strains was uploaded under the same BioProject as the data in Crombie *et al.* 2019. We have updated the sentence that describes our dataset in the first results section to make this clearer (below).

"To explore the genetic diversity in the *C. elegans* species, we examined whole-genome sequence data from 609 wild *C. elegans* strains isolated from six continents and several oceanic islands, including 103 wild strains that have not been studied previously."

The argument for a Pacific origin is a little uncertain in my opinion. First, it would good to define the geographic region meant by "Pacific" clearly. For example, the authors might mean the western coasts of the Americas, the Pacific Islands, and/or eastern and southeastern Asia. Second, Hawaii is an international tourist destination, has extensive importation of crop species, and has a population consisting of with many immigrant peoples. High variation may simply reflect immigration of C. elegans populations from places not sampled by the authors. It is my opinion that this point should be mentioned along with the authors very limited sampling in Asia.

We agree with the reviewer that there are alternative explanations for the high diversity in the Pacific islands/coast. We have updated this section of results to define more clearly what we mean by Pacific region ("Pacific region that encompasses the Hawaiian islands, New Zealand, and the Pacific coast of the United States"). We have also added a sentence that discusses alternative explanations and that acknowledges our limited sampling in Asia.

Added in MS: "The current geographic location with the most divergent strains might not reflect the origin of the species because worldwide sampling is uneven and some sites, including Asia, have not been sampled extensively. Therefore, *C. elegans* could have originated elsewhere and spread throughout the Pacific region. "

If the authors decide to keep the early focus on population structure, it would be informative to understand the global distribution of isotypes. Were they always found in the same location?

We have chosen to keep the early focus on population structure because we believe it provides an important context for the results that follow (for example, finding that hyper-divergent haplotypes are often shared between divergent genetic groups). We have added additional detail about the global distribution of isotypes.

Added in MS: "The majority of wild strains (368 strains) with the same isotype (87 isotypes) were sampled from the same location, which is consistent with previous observations that local habitats frequently consist of clonal populations[15,29]. However, we discovered 14 isotypes that were sampled from locations at least 50 km apart (Supplementary Fig. 1a, Supplementary Table 2), suggesting that individuals can migrate long distances in the wild."

Given the likely extended LD in C. elegans, the authors should explore downsampling their variant datasets for PCA to limit overestimation population divergence based on dependent, linked polymorphism. In addition, the authors should provide variance explained on each axis.

We thank the reviewer for this concern. We downsampled the variant dataset using three LD thresholds, 0.2, 0.6, and 0.8, and presented all results in Fig. 1c-f and Supplementary Fig. 2,3. We also provided variance explained of each axis for all these PCA plots. The additional pruning did not significantly alter the conclusion that the species is likely divided in three distinct genetic groups.

The authors assert that the several major sweeps in the C. elegans genome that they previously identified may have occurred during the out of the Pacific event. I think that Line 102-103 should read Extended Data Fig 2. It would be valuable to see all of the chromosomes rather than just these selected chromosomes for comparison sake. Also the beeswarm plot, presented by chromosome would be a helpful addition.

We have fixed the error. We have updated Supplementary Fig. 4a to show all six chromosomes, but have opted to change the x-axis label from “Fraction swept” to “Fraction of most frequent haplotype” as previous work has shown that chromosomes II and III have not undergone selective sweeps. We present beeswarm plots in Supplementary Fig. 4b.

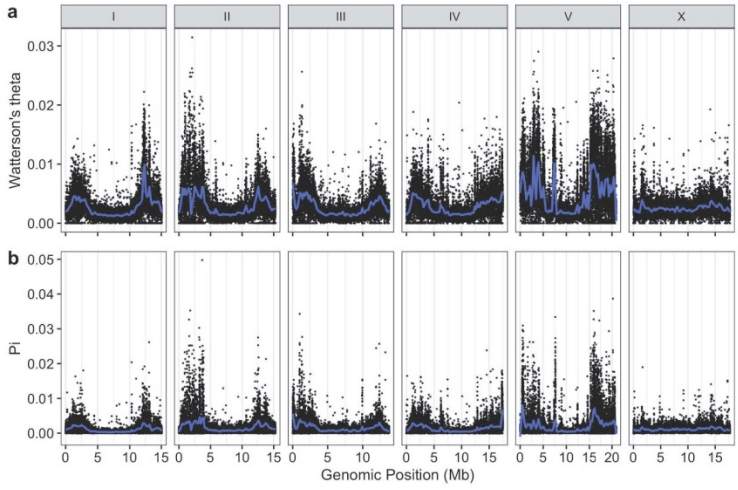
The authors move on to the major focus of the manuscript, the role of balancing selection in determining genomic diversity in genome of C. elegans. They begin by observing regions with exceptionally high values of diversity using standard estimators. They refer to these as high divergence, but I think using statistics like theta and pi it is better to use the term diversity.

We agree with the reviewer and we have replaced ‘divergence’ with ‘diversity’ in this section of the results.

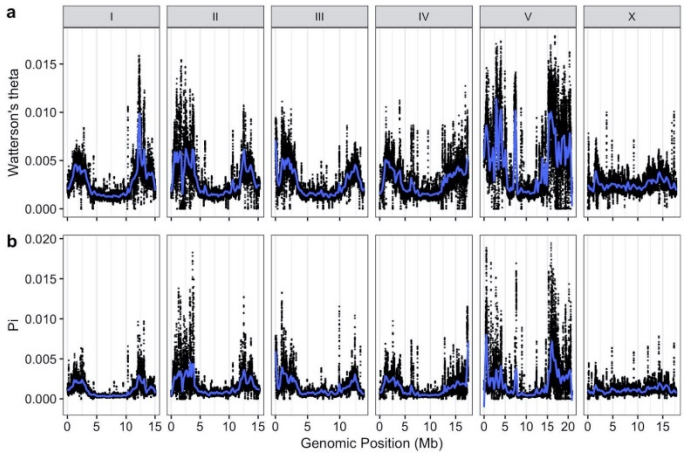
It should be noted that the authors use very small window sizes in these scans (1kb). What is the length of LD in C. elegans and does this make sense? Did the authors screen windows with only small numbers of accessible bases?

For consistency, we performed the diversity and Tajima’s *D* scans using the same 1 kb intervals we used to classify hyper-divergent regions. However, to address the reviewers concerns regarding LD and accessible bases, we also calculated the diversity statistics and Tajima’s *D* using non-overlapping 10 kb windows and 10 kb windows with 1 kb step. These results are displayed below and did not change the conclusions we made from the 1 kb analysis.

(1 kb window 1 kb step size)



(10 kb window 1 kb step size)



Supplementary figure 4 is listed as evidence of poor alignment, but perhaps it would be better to look at alignment statistics in these windows directly rather than showing two examples.

Supplementary Fig. 9 and Supplementary Table 6 in the submitted manuscript provide this information. Supplementary Fig. 9 shows that hyper-divergent variants have less coverage and more variants than non-divergent regions and Supplementary Table 6 provides a numeric summary of these data.

The authors next use a set of empirical thresholds to define hyper divergent regions with whole genome assemblies and their illumina data. They describe a procedure by which they chose their empirical thresholds, but do not provide any data supporting their use of their thresholds or any of the results from their optimization. The authors should provide a clear justification for the threshold using the datasets from different thresholds and by showing where these particular thresholds fall in the overall distributions of diversity statistics/alignment statistics. How did the authors figure out what is a true divergent region to use to assess false positives and negatives?

As the reviewer alludes to, it is not possible for us to know the hyper-divergent "truth" set and therefore assess false positive and negative rates. To circumvent this limitation, we leveraged the 14 long-read assemblies we generated and the previously assembled genomes of CB4856 and N2 to assess the parameters used to call hyper-divergent regions for the remaining *C. elegans* strains for which we only had short-read sequencing data. We realize this procedure was not adequately summarized in the methods section of the submitted manuscript. We added the necessary details to the "Characterization of hyper-divergent regions" section of the Methods as below. A summary of the parameter optimization procedure can be found in Supplementary Fig. 7.

Updated method description:

Characterization of hyper-divergent regions

To characterize hyper-divergent regions across the *C. elegans* species, we first analyzed short-read and long-read alignments of 15 isotypes. For all non-overlapping 1 kb windows in the reference *C. elegans* genome, we calculated the number of small variants (SNVs and indels) (variant count) using the coverage subroutine in the BEDtools (v2.27.1)[97] suite and the average sequencing depth using mosdepth (v0.2.3)[98]. We converted the coverage values to coverage fraction (average sequencing depth of the window/genome-wide average depth). In parallel, we aligned all 14 long-read assemblies we generated along with a long-read assembly for the Hawaiian isotype CB4856[28] to the N2 reference genome (WS255)[99]

using NUCmer (version 3.1) with the following parameters: *--maxgap=500 --mincluster=100*)[100]. Coordinates and identities of the aligned sequences were extracted from the alignment files using the 'show-coords' function with NUCmer. Then, we calculated the average alignment coverage (alignment coverage) and average alignment identity (alignment identity) for each non-overlapping window in the reference genome. Next, we used the long- and short- read alignment datasets to identify the optimal coverage fraction and variant count parameters to apply to the rest of the population for which we do not have long-read sequence data. We tested a wide range of parameters to define hyper-divergent regions from short-read and long-read alignments. For the short-read based approach, we classified each window as hyper-divergent if its variant count ≥ x or coverage fraction < y% or both; we also classified windows that are flanked by two hyper-divergent windows as hyper-divergent. Then we clustered contiguous hyper-divergent windows and defined clusters that are greater than or equal to 9 kb of N2 reference genome length as hyper-divergent regions[27]. For the long-read based approach, we classified each window as hyper-divergent if its alignment coverage < z% or alignment coverage < w% or both; we also classified windows that are flanked by hyper-divergent windows as hyper-divergent. Then, we clustered contiguous hyper-divergent windows and defined clusters that are greater than or equal to 9 kb of the N2 reference genome length as hyper-divergent regions. Because we lacked a "true" hyper-divergent region dataset to which we could tune our parameters, we identified the set of x, y, z, w values that maximized the overlap between hyper-divergent regions identified by short- and long-read based approaches (Supplementary Fig. 8). To minimize the amount of false positives that we detected, we manually validated hundreds of regions that were classified as hyper-divergent using IGV[101]. Using this optimization, we selected the optimal short-read classification parameters (variant counts ≥ 16 and coverage fraction < 35%), which we then applied to all 327 non-reference isotypes. With these classification parameters, we identified a similar size of hyper-divergent regions (3.2 Mb) in CB4856 to the total size of hyper-divergent regions (2.8 Mb) identified in CB4856 previously[27]. Additionally, we confirmed that selected parameters do not detect any hyper-divergent region from short-read alignments of N2 reference strain to its own reference genome. To exclude large deletions that could be classified as hyper-divergent regions, we filtered out hyper-divergent regions without any window with high variant density that exceed variant count threshold.

The authors identify 20% of the genome that is hyper-divergent from the reference genome in at least one accession. This use of reference polarization makes some of the subsequent analyses difficult to interpret. For example, I am not sure why it makes sense to calculate the fraction of the genome that is hyperdivergent from the reference in each line. Does this not simply reflect the relationship to the reference line? Perhaps the authors intend to show which regions harbor the highest number of rare divergent alleles?

We agree with the reviewer that reference polarization plays a role in the patterns shown in Fig. 3. However, we feel justified in presenting results in this way because the N2 is a widely used reference strain and we have no other way to show how frequently the same regions are classified as hyper-divergent in wild genomes (which was our aim in Fig. 3a). However, we agree with the reviewer that there are several places in the submitted manuscript where we do not make it clear that our analyses are reference-biased. We have therefore changed the axis and legends of figures that used the term 'Genomic position' to 'Reference genome position', to make it clearer that this is relative to the N2 reference genome.

The divergent alleles identified by the authors are very interesting. There are a number of possible explanations for such alleles including random chance, introgression and balancing selection. The authors immediately suggest balancing selection as their preferred hypothesis, but fail to test any other hypothesis.

We agree with the reviewer that we should discuss all potential origins for the divergent haplotypes. Although we observed strong signatures of balancing selection and identified hyper-divergent haplotypes in these regions, we acknowledge that such hyper-divergent haplotypes could have originated from the ancestral population or introgression from other species or both. Unfortunately, we cannot directly test whether they have originated via retention of ancestral diversity (i.e. by looking for trans-specific polymorphisms) or via introgression because a closely related sister species has not yet been identified for *C. elegans*. The observation that 20% of the genome is hyper-divergent and enriched for environmental response genes likely excludes the possibility that all of these regions have been kept in the *C. elegans* population by chance. We now provide additional evidence for balancing selection (see below). We have

toned down the conclusion of retained ancestral genetic diversity by expanding the discussion of the possibility that adaptive introgressions have been maintained over the evolutionary history of the species. Additionally, we have changed the title accordingly.

The evidence they present is an increased Tajima's D at these loci. While this result is consistent with their hypothesis, Tajima's D is sensitive to many demographic factors and should be supported with additional tests.

We agree with the reviewer that Tajima's *D* can be influenced by demographic factors. However, we did provide additional evidence in the submitted manuscript. We showed in Fig. 5 and Supplementary Fig. 15-17 that the gene trees in hyper-divergent regions are discordant with genetic groups inferred using PCA. Furthermore, we presented evidence that these regions are enriched for genes that are typically maintained by balancing selection. We also noted that our scans identified genomic regions that were previously shown to be maintained by balancing selection. However, to provide additional support, we have now calculated and present the β statistic, a summary statistic that has been shown to be less influenced by demography than Tajima's *D* (updated Fig. 4). In the revised manuscript, we demonstrated that genomic regions that are frequently classified as hyper-divergent across the species exhibit high values for both statistics (Fig. 4c-f), which supports our hypothesis that these regions are under long-term balancing selection. However, we note in the revision that levels of diversity are underestimated when calculated using short-read sequence data, influencing both Tajima's *D* and β (Supplementary Fig. 10).

Also, large sections of the genome are argued by the authors to have been subject to directional selection, so wouldn't neutrally evolving loci also show elevated genomic Tajima's D if preselected to carry some genomic variation (as is the case here)?

We agree with the reviewer that pervasive directional selection (selective sweeps) can lower Tajima's *D*, therefore neutrally evolving loci would show elevated Tajima's *D* relative to the swept genome. Instead of previous comparisons for average Tajima's *D* values, we analyzed the enrichment of genomic bins or genetic variants with very high values of both Tajima's *D* and β (Fig. 4e-f), which is less likely to happen at neutrally evolving loci, and found strong evidence of balancing selection.

Also, how do the authors think that increased LD fits with their hypothesis of balancing selection? Long LD can occur at balanced loci under some models, but certainly not all, so perhaps the authors should provide some discussion of this.

We agree with the reviewer that not all models of balancing selection have increased LD. We edited the text to remove the suggestion that increased LD in hyper-divergent regions is evidence of balancing selection (below).

Added in the MS:
"We found that these regions range from being divergent in a single isotype to divergent in 280 isotypes (85%). Interestingly, we find that SNVs in hyper-divergent regions have a lower rate of linkage disequilibrium (LD) decay than SNVs within non-divergent regions on the autosomal arms (Fig. 4b), suggesting that these regions are inherited as large haplotype blocks. We performed genome-wide scans using commonly used statistics (Tajima's *D* and standardized β) to identify regions under long-term balancing selection[22,24,36], which could explain the presence of hyper-divergent haplotype blocks that are shared by a substantial fraction of isotypes[19]."

In the gwas, why was the overlap calculated post hoc rather than using the genotype of each line in the divergent region in the gwas? Obviously, kinship/relatedness should still be accounted for in such an analysis.

In response to Reviewer #2's points, we removed the GWAS analysis from the manuscript.

The authors do not show that these alleles are ancient (which sets up the claim of long term balancing selection). The argument for these alleles being ancient is that they are highly diverged, and that these divergences are similar to those observed between species in a related lineage. I think this result is suggestive (assuming that the error rates are the same in these regions as in other regions), but this can not be seen as a formal dating of the alleles as implied by the term ancient. There approach does not exclude variation in recombination rates, mutation rates, or introgression as contributing factors. If the authors identify some of these alleles in related species, that would provide evidence for ancient balancing selection. Otherwise, I suggest that the authors instead argue for balancing selection, and place the high divergences in context of previous studies that could compare alleles across species.

We agree with the reviewer that, because we have not shown conclusively that the hyper-divergent alleles represent retained ancestral genetic diversity, we cannot say for sure that they are ancient. As suggested by the reviewer, we now present additional evidence to show that these haplotypes are maintained by balancing selection. Furthermore, we have updated our discussion section to provide a more balanced discussion of the potential origins of these divergent haplotypes. We have changed the title of our manuscript accordingly.

Reviewer 2

This is an extremely well written MS that will be of great general value to the broader scientific community. This MS describes tracks of old polymorphism within C. elegans sampled world-wide, an argues that these tracks of high genetic diversity have been maintained by balancing selection. I have a few suggestions to the authors that I believe will substantially improve the plausibility and interpretability of their results.

1. Throughout the manuscript the authors use the phrase "haplotypes" to both describe genetically unique isolates and to describe patterns of genetic diversity at specific loci. As someone who doesn't think about C. elegans all the time, I had to remember that these wild isolates are basically inbred due to selfing. It took me a while to recognize that I was conflating my use of the word haplotype (which I typically think of as being restricted to a specific region of the genome) and the way the authors were using it (at least in some places). I think that the first confusing instance of the phrase haplotype is on line 78 ("We identified 328 distinct genome-wide haplotypes (henceforth, referred to as isotypes)"). The next instance of haplotype is the genetically localized on on line 103. It is a minor detail but I think that making the language more clear would make the paper more accessible to the outside reader.

We agree that this is confusing and have updated the text to not refer to strains as haplotypes.

2. I am skeptical of the GWAS and the enrichment test that the authors did. I recognize that the GWAS approach that the authors performed follows an established pipeline that is designed to greatly reduce the multiple testing burden. While I think that the approach is reasonable, I have a set of suggestions: First, the enrichment test you performed and reported as the result of a hypergeometric test is problematic for a number of reasons: (1) it is unclear if the test is driven by one or more than one of the GWAS results (there were two or three phenotypes); (2) the hypergeometric test might not be the most appropriate given that the assumption of independence is violated with correlated signal across the genome - you could use any number of the overlap type tests that have been designed for genomic regions; (3) no actual results of the GWAS are given (like regions, p-values, etc, etc) and all we know is that the results beat a non-conservative implementation of Bonferroni. To solve issues #1 and #2, I think that the authors need to implement a permutation test where they permute the phenotype data, repeat the GWAS, and the enrichment test. If their actual data beat the permutation in terms of enrichment, then I will believe the result. If the GWAS do not beat the permutation then I will think that the result is artifactual and driven by an inappropriate use of multiple testing correction, enrichment test, or both. Also, the authors should report the results of their GWAS rather than point the reader to an technical supplemental table that I couldn't really understand (there are more than three columns/phenotypes, why?)

We understand the reviewers point that the hypergeometric test might not be the best test for enrichment. We re-performed the analysis using a permutation-based genomics method, implemented in the regionR R package and found the overlap to be marginally significant. However, in an unrelated project, we found that using an imputed variant set to perform GWA mappings in *C. elegans* causes many spurious associations, particularly in hyper-divergent regions. These false QTL are likely caused by imputation inferring variants in hyper-divergent regions inaccurately. Therefore, we performed the GWA analysis again using a filtered but not imputed variant set. This analysis greatly reduced the number of significant associations we identified, which were most likely false positives. Accordingly, no significant overlap was identified between the identified QTL and hyper-divergent regions. We removed this analysis from the manuscript and thank the reviewer for the suggestions.

3. I really want to know what all those extra genes in the hyper-divergent are. Are they paralogs of other neighboring genes? For instance, are we observing the "accordion" genome effect where organisms duplicate and diversify genes to accommodate different functions? Or, are they molecularly unique? Any indication would be super informative and is conspicuously absent from a discussion.

We agree with the reviewer that this is important information that was lacking in the submitted version of the manuscript. For the three (Fig. 5) and six/seven haplotype regions (Supplementary Fig. 17), we have functionally annotated all non-conserved genes using InterProScan and searched the protein sequences against the N2 proteome using BLASTP. We did not perform this analysis on the *peel-1 zeel-1* locus (Supplementary Fig. 16) as the only two unconserved genes are the toxin/antitoxin elements. In the three haplotype region, the majority (13/16) of the non-conserved genes have homology to other genes in the region (*srx-101*, *F40H7.12*, or *F19B10.10*, all of which are within this region in the reference haplotype). The remaining loci include genes with homology to genes elsewhere in the N2 reference genome. Of the 25 genes in the six/seven haplotype region that are not conserved across all haplotypes, ten are alleles of the reference haplotype (N2) genes. The remaining 15 do not have a clear one-to-one relationship with a gene in the reference haplotype. Seven of these 15 have homology to *F54E12.2* (present in the reference haplotype). Six have homology to either *M04C3.1*, *F19B2.5*, or *F54E12.2*, all of which are genes with SNF2 family N-terminal domains and which exist elsewhere in the N2 reference genome. Of the remaining two genes, one has homology to Y113G7B.15, which is present in the reference haplotype, and the other has homology to W09C3.8, a gene on chromosome I in the reference genome. Given these data, it appears that many of the haplotype-specific genes in these regions can be explained by the "accordion" effect of duplication and diversification as suggested by the reviewer. We have updated our results section, the figure legends, and added an additional supplementary table (Supplementary Data 4) to present these results. Furthermore, when performing these extra analyses, we identified a small number of cases where the predicted genes did not have clear homology to an N2 gene or a Pfam domain, and were likely spurious predictions. We have removed these from our updated plots.

Reviewer 3

In this report, Lee et al use a combination of resequencing, GWAS on fitness-related traits to growth on different natural bacterial food sources, and analysis of previously published transcriptomics to characterize balancing selection in wild strains of C. elegans. Perhaps due to its mating system, study of C. elegans has led to the publication of a number of impactful reports characterizing the effect of various evolutionary forces on segregating genetic variation (e.g. Rockman et al, 2010 Science, Andersen et al, 2012 Nature Genetics). Here, the authors use short and long-read resequencing to study the effects of balancing selection in wild populations of C. elegans.

This work builds on a number of previous reports that has implicated the importance of balancing selection in maintaining genetic variation in C. elegans. In three previous reports, specific loci that affect responses to natural products – avermectins and pheromones – were identified (Ghosh et al 2012 Science, Greene et al 2016 Nature, Lee et al, 2019 Nature E&E). Each of these loci showed signatures of balancing selection using Tajima's D and two of these loci showed a remarkable amount of nearby genetic variation consistent with the action of long-term balancing selection. Two additional reports have identified toxin-antitoxin elements that show signatures of balancing selection, although it is unknown if these toxin-antitoxin elements are under balancing selection or linked to loci that are (Seidel, 2008 Science, Ben-David et al 2017 Science). Finally, resequencing of a C. elegans wild strain identified many remarkably divergent regions throughout the genome (Thompson et al 2015, Genetics). The high amount of variation is consistent with balancing selection retaining ancient genetic variation conferring functional changes to nearby genes, similar to human haplotypes in the MHC locus.

Lee et al build upon these results using resequencing of wild strains of C. elegans using a combination of short-read resequencing (609 strains) and long-read sequencing (14 strains). From this data, they identify 366 new hyper divergent regions. Are these hyper divergent regions under balancing selection? They address this in three ways. 1) They use population genetics to show these regions have higher Tajiama's D than the rest of the genome 2) They show these regions are enriched for genes that are involved in environmental sensation, consistent with environmental heterogeneity. 3) They use GWAS and transcriptomics to characterize the response to natural bacteria and pathogens and show they overlap with these regions.

Overall, this paper should be of interest to evolutionary biologists – broadly to those interested in the role of balancing selection in maintaining genetic diversity in a population and the effect of mating system on evolutionary trajectories. It should also be of interest to the study of C. elegans evolutionary biology.

Some specific comments on the manuscript:

1. I'm not sure if this is intentional but there is no Introduction section. This would probably be useful to more formally introduce their working model for C. elegans evolutionary trajectory that frames the work. 1. Male/Female species in the Hawaiian islands with genetic diversity on chromosomal arms shaped by background selection. 2. Evolution to androdioecy followed by loss of genetic diversity except for regions under balancing selection. 3) Recent spread of C. elegans across the globe accompanied by selective sweeps across multiple chromosomes. It might be useful to include a figure panel that summarizes this model. However, this is an editorial decision for the authors to make.

We thank the reviewer for this concern. We have added an introduction section to our manuscript that describes the evolution of selfing in *Caenorhabditis* and effects on genetic diversity, the species- and genome-wide patterns of genetic diversity in *C. elegans*, and previous evidence of ancient diversity being maintained by balancing selection.

2. Figure 1 – It is difficult to distinguish the blue colors for Hawaiian vs Atlantic strains.

We updated the color palette for Fig. 1 and all subsequent figures that used the same color palette.

3. Figure 2c – Are the authors underestimating the amount of genetic diversity that is contained within these regions as short read sequences are unable to align fully in these regions? Further, why are there two clusters of strains? Is there a reason not to follow their previous color coding by geographical origin?

We are indeed underestimating the diversity contained within these regions using short reads, as discussed and shown by the long-read based analysis. The two clusters of strains correlate with the population structure inferred using PCA. We have updated Fig. 3c so that the points are now colored by genetic group. We also added a statement for the underestimation of diversity in hyper-divergent regions.

Added in MS:
"The majority of these regions (69%) are found on autosomal arms and contain 10.3-fold higher variant (SNVs/indels) densities than the non-divergent autosomal arm regions (16.6-fold more than the genome-wide average) (Supplementary Table 6, Supplementary Fig. 9). Although variant counts are likely underestimated because short sequencing reads often fail to align to the reference genome in hyper-divergent regions, we note that the level of genetic diversity in these regions is similar to the level of genetic diversity reported in outcrossing *Caenorhabditis* species[34,35]."

4. Figure 2d – Why is there more genetic diversity in Atlantic populations? Is this consistent with the out of the Pacific origin for C. elegans? The authors indicate that these strains show less of the swept haplotypes. Would that be enough to explain the increased diversity?

As the reviewer suggests, the higher levels of genetic diversity in the Atlantic isotypes is because they have not undergone the same selective sweeps that their continental counterparts have. We have added a sentence to the results section to explain this point.

Added in MS:
"In addition to the Hawaiian isotypes that were reported to have avoided these selective sweeps[31], we found that the genomes of isotypes from Atlantic islands (*e.g.* Azores, Madeira, and São Tomé), in contrast to continental isotypes, show less evidence of the globally distributed haplotype that swept through the species (Supplementary Fig. 4, Supplementary Table 3)."

5. I found it very confusing how the hyper divergent regions were defined. I understand parameters were chosen to maximizing overlap between long and short read technologies but it was unclear where the original definition is. Can the authors justify more what defines a hyper divergent region?

As per our response to reviewer 1, we realized that we did not include sufficient detail in the original submission about the parameter optimization procedure and selected parameters for defining hyper-divergent regions. We have moved Extended Data Fig. 4 to Fig. 2 to demonstrate how we defined the hyper-divergent regions. We also have updated the Methods section to reflect all of the details and a summary of the parameter optimization procedure can be found in Supplementary Fig. 7.

Updated method description:
Characterization of hyper-divergent regions
To characterize hyper-divergent regions across the *C. elegans* species, we first analyzed short-read and long-read alignments of 15 isotypes. For all non-overlapping 1 kb windows in the reference *C. elegans* genome, we calculated the number of small variants (SNVs and indels) (variant count) using the coverage subroutine in the BEDtools (v2.27.1)[97] suite and the average sequencing depth using mosdepth (v0.2.3)[98]. We converted the coverage values to coverage fraction (average sequencing depth of the

window/genome-wide average depth). In parallel, we aligned all 14 long-read assemblies we generated along with a long-read assembly for the Hawaiian isotype CB4856[28] to the N2 reference genome (WS255)[99] using NUCmer (version 3.1) with the following parameters: *--maxgap=500 --mincluster=100*)[100]. Coordinates and identities of the aligned sequences were extracted from the alignment files using the 'show-coords' function with NUCmer. Then, we calculated the average alignment coverage (alignment coverage) and average alignment identity (alignment identity) for each non-overlapping window in the reference genome. Next, we used the long- and short- read alignment datasets to identify the optimal coverage fraction and variant count parameters to apply to the rest of the population for which we do not have long-read sequence data. We tested a wide range of parameters to define hyper-divergent regions from short-read and long-read alignments. For the short-read based approach, we classified each window as hyper-divergent if its variant count ≥ x or coverage fraction < y% or both; we also classified windows that are flanked by two hyper-divergent windows as hyper-divergent. Then we clustered contiguous hyper-divergent windows and defined clusters that are greater than or equal to 9 kb of N2 reference genome length as hyper-divergent regions[27]. For the long-read based approach, we classified each window as hyper-divergent if its alignment coverage < z% or alignment coverage < w% or both; we also classified windows that are flanked by hyper-divergent windows as hyper-divergent. Then, we clustered contiguous hyper-divergent windows and defined clusters that are greater than or equal to 9 kb of the N2 reference genome length as hyper-divergent regions. Because we lacked a "true" hyper-divergent region dataset to which we could tune our parameters, we identified the set of x, y, z, w values that maximized the overlap between hyper-divergent regions identified by short- and long-read based approaches (Supplementary Fig. 8). To minimize the amount of false positives that we detected, we manually validated hundreds of regions that were classified as hyper-divergent using IGV[101]. Using this optimization, we selected the optimal short-read classification parameters (variant counts ≥ 16 and coverage fraction < 35%), which we then applied to all 327 non-reference isotypes. With these classification parameters, we identified a similar size of hyper-divergent regions (3.2 Mb) in CB4856 to the total size of hyper-divergent regions (2.8 Mb) identified in CB4856 previously[27]. Additionally, we confirmed that selected parameters do not detect any hyper-divergent region from short-read alignments of N2 reference strain to its own reference genome. To exclude large deletions that could be classified as hyper-divergent regions, we filtered out hyper-divergent regions without any window with high variant density that exceed variant count threshold.

6. I think it is important to summarize more of the results from the long-read sequencing. How similar are the divergent regions produced by the long-read sequencing vs. small-read sequencing. How many large deletions/insertions are missed?

We have updated our manuscript to provide more detail on characterizing hyper-divergent regions, which involves using the long-read assemblies to optimize the short-read based classification approach. Results from comparisons between the long-read and short-read based approaches can be found in Supplementary Fig. 7. On average, 60-70% hyper-divergent regions overlap between short-read and long-read based classifications.

How do they affect gene function?

We agree it would be fascinating to understand the functional differences between different divergent haplotypes, but that would require extensive future laboratory studies. However, some clues can be found in previously characterised divergent regions. For example, at the *roam-1* locus, two divergent haplotypes underlie two density-dependant foraging strategies. We describe this work (and additional studies) in the revised Discussion section.

How many of the missing/added genes have homologs in nematodes and other species. Do any have homology to known toxin-antitoxin genes?

As per our response to reviewer 2, we agree that this is important information that was lacking in the submitted version of the manuscript. We added data (Supplementary Data 4) that provide functional annotations for all non-conserved genes in the three regions we analyzed in the manuscript. We have opted against performing a genome-wide analysis of all non-conserved genes because each region

requires extensive manual curation to ensure that "differences" in gene content are not simply differences/artefacts arising from automated gene prediction. However, to address the reviewers second question about toxin/antidote systems, we used BLASTP to search the predicted protein sets of all 16 isotypes with long-read assemblies for sequences with homology to previously identified toxin/antidote genes (*peel-1*, *zeel-1*, *sup-35*, and *pha-1*). We looked at hits in wild isotypes that were better (*i.e.* hits that had a lower e-value) than the top non-query N2 hit (as those hits below this were likely alleles of this non-query N2 gene). In all searches other than *sup-35*, the hits were found only in those isotypes that shared the N2 haplotype and were therefore likely to represent alleles of *peel-1*, *zeel-1*, or *pha-1*, rather than additional, uncharacterised toxin/antidote systems. In the *sup-35* search, all wild isotypes had between two and three hits above the top non-query N2 hit. Although this observation is potentially interesting, understanding the relationship of these genes to *sup-35* would require manual curation of these regions and further analysis (orthology, gene trees, etc). Moreover, proving that these loci represent novel toxin/antidote systems is not possible without laboratory work. As a result, we have opted not to include this analysis in the manuscript, but we agree with the reviewer that this is an interesting question and one that should be addressed in future studies.

Decision Letter, first revision:

Our ref: NATECOLEVOL-200911636A

17th February 2021

Dear Dr. Andersen,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Ecology & Evolution manuscript, "Balancing selection maintains hyper-divergent haplotypes in C. elegans" (NATECOLEVOL-200911636A). Please carefully follow the step-by-step instructions provided in the personalised checklist attached, to ensure that your revised manuscript can be swiftly handed over to our production team.

Please get in contact with us immediately if you anticipate it taking more than two weeks to submit these revised files.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

In recognition of the time and expertise our reviewers provide to Nature Ecology & Evolution's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Balancing selection maintains hyper-divergent haplotypes in C. elegans". For those reviewers who give their assent, we will be publishing their names alongside the published article.

Nature Ecology & Evolution offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

Cover suggestions

As you prepare your final files we encourage you to consider whether you have any images or

illustrations that may be appropriate for use on the cover of Nature Ecology & Evolution.

Covers should be both aesthetically appealing and scientifically relevant, and should be supplied at the best quality available. Due to the prominence of these images, we do not generally select images featuring faces, children, text, graphs, schematic drawings, or collages on our covers.

We accept TIFF, JPEG, PNG or PSD file formats (a layered PSD file would be ideal), and the image should be at least 300ppi resolution (preferably 600-1200 ppi), in CMYK colour mode.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

Nature Ecology & Evolution has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article.

Please note that you will not receive your proofs until the publishing agreement has been received through our system.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:

[REDACTED]

If you have any further questions, please feel free to contact me.

[REDACTED]

Reviewer #1:
Remarks to the Author:
In general, the authors have made a good faith effort to address my comments on the previous draft, and I continue to think the study is very interesting and appropriate for this journal. I alsothink that it would be preferable if the authors could provide some simulation based data to show that the high diversity in these regions can not be explained by a random retention of ancestral variation in some

parts of the genome (perhaps combined with historical balancing selection prior to the species wide loss of genetic diversity). Still, I the authors make a strong argument that balancing selection is likely an important factor shaping diversity in this system.

Reviewer #2:
Remarks to the Author:
My comments have been adequately addressed and I have no further critique. I think that this paper will be an important contribution to the community.

Reviewer #3:
Remarks to the Author:
The authors have responded to all my points. I recommend for this to be published in Nature E&E

Final Decision Letter:

26th February 2021

Dear Erik,

We are pleased to inform you that your Article entitled "Balancing selection maintains hyper-divergent haplotypes in C. elegans", has now been accepted for publication in Nature Ecology & Evolution.

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

The subeditor may send you the edited text for your approval. Once your manuscript is typeset you will receive a link to your electronic proof via email, with a request to make any corrections within 48 hours. If you have queries at any point during the production process then please contact the production team at rjsproduction@springernature.com. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see www.nature.com/authors/policies/index.html). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

Nature Ecology & Evolution is a Transformative journal and offers an immediate open access option through payment of an article-processing charge (APC) for papers submitted after 1 January, 2021 . In the event that authors choose to publish under the subscription model, Nature Research allows authors to self-archive the accepted manuscript (the version post-peer review, but prior to copy-editing and typesetting) on their own personal website and/or in an institutional or funder repository where it can be made publicly accessible 6 months after first publication, in accordance with our self-archiving policy. <a href=""https://www.nature.com/nature-research/editorial-policies/self-archiving-

and-license-to-publish"">Please review our self-archving policy for more information.

Several funders require deposition the accepted manuscript (AM) to PubMed Central or Europe PubMed Central. To enable compliance with these requirements, Nature Research therefore offers a free manuscript deposition service for original research papers supported by a number of PMC/EPMC participating funders. If you do not choose to publish immediate open access, we can deposit the accepted manuscript in PMC/Europe PMC on your behalf, if you authorise us to do so.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Ecology & Evolution as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Ecology & Evolution logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

You can generate the link yourself when you receive your article DOI by entering it here: <a href="http://authors.springernature.com/share">http://authors.springernature.com/share<a>.

Yours sincerely,

[REDACTED]

P.S. Click on the following link if you would like to recommend Nature Ecology & Evolution to your librarian http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**