Dear Dr Fumagalli and Reviewers,

We appreciate your engagement in reviewing our preprint and constructive feedback. Please find our replies to the comments below. We hope the revision has dealt with the points raised for improvements. The referred line numbers in our replies correspond to the PDF on bioRxiv (doi: https://doi.org/10.1101/2021.12.22.473882) instead of the track change file on PCI. Our replies contain three figures, which we believe are too specific for the reviewers' comments to be included even in the supplementary figures, but we are happy to transfer them to the supplementary materials if the recommender and reviewers find that they are essential for the manuscript. We look forward to further correspondence.

Best regards,
Jun Ishigohoka

Decision for round #1 : *Revision needed*

Minor revision

The preprint has been reviewed by two experts in the field. They both found the study of merit and suggested several points to modify, mostly in the presentation of the methods. There are also several additional analyses which are suggested to do. While not all of them would add significantly to the study, I would encourage the Authors to at least attempt to reply to each point raised.

*by **Matteo Fumagalli**, 20 Dec 2023 16:03*
Manuscript: ***https://doi.org/10.1101/2021.12.22.473882***
version: 2

Review by Claire Merot, 24 Nov 2023 15:01

Principal component analysis (PCA) are increasingly used to capture and understand the distribution of genetic variation accross many samples and along the genome. In fact, PCA can also be performed on windows along the chromosomes, a method called local PCA (Ralph and Li, 2019) that put in evidence discrepancies in the structure of genetic variation. It is most frequently used to detect non-recombining haploblocks typically induced by chromosomal rearrangements but may also reveal any long block of linked loci whether this is due to reduced recombination, selection, low migration, etc.

The present article provides a very relevant and in-depth exploration of how the recombination landscape may affect local PCA patterns both in empirical data (Blackcap) and simulated data. The results highlight that low recombination on its own may explain outliers windows in which PCA patterns show consistency among several adjacent windows. While this results may be expected, it nicely complement other previous exploration of the methods that did not clearly distinguished cases with and without linked selection.

The overall result makes sense given how important LD is in driving PCA patterns. The methods used to explore the data and infer results are sound and well-explained. Actually, reading the methods section, we discover that much more has been done that what is visible in the results. In fact, beyond the methodological emphasis, the paper also provides a thorough exploration of patterns of genetic variation along the genome of blackcap, demonstrating the presence of at least one or two polymorphic inversions, as well as the geographic structure of the species. I do believe the results could be a little

more complete to include a few information on the new reference genome, the strong work done to confirm the inversion on chromosome 12, and the in-depth exploration of the simulations.

The discussion is slightly long but really well-written. It explains very well tricky concepts such as genealogy, recombination impact, etc…It serves well the purpose to understand the subtilty of the simulation results. The figures are beautiful and clear. I particularly appreciated the schematic conceptualisation. Supplementary materials is dense and reflects how much work has gone in each sub-part of this paper. It is thus even more impressive to end up with a very clear manuscript unified in a single message.

Overall, this article was a pleasure to read, is relevant for current research in evolutionary genomics, and I have very few comments.

Thank you for inviting me to review

Claire Mérot

· Major comments :

1- Important informations are missing. What is the size of the clusters of windows? How does such parameter, that will also depend on the density of SNPs may impact the results ?

Reply:

Thank you for mentioning the lack of clarity and rationale on window size applied. The window size we chose for the analyses using lostruct is 1,000 SNPs-long.We state this information in the Materials & Methods (revision 1 P28 L553) as well as in Results (revision 1 P7 L127). We defined the window size based on the number of SNPs instead of physical length (bp) to make sure that PCA is carried out on the same amount of polymorphisms and to exclude the possibility that outlier windows are due to the lack of data. Our chosen value of 1,000 SNPs was set longer than the default value of lostruct (100 SNPs) to reduce computational demand, especially in long chromosomes, and to increase the signal/noise ratio. By further increasing the window size, short regions with distinct patterns of genetic variation may be missed, because such regions may not cover a sufficiently high number of windows to pass our threshold for calling outlier regions (i.e. at least five outlier windows). Conversely, windows with fewer SNPs may discover additional outlier regions, which were too short to be identified with our current criteria at the expense of more false positives due to increased noise.

2- The introduction and review of existing litterature tends to be caricatured. There is no need to claim that low-recombination alone was not considered before to make this study relevant. On the contrary, most localPCA studies have uncovered inversions and low-recombining regions (with or without support for additional selection). Several papers have considered the impact of low-recombination on genetic statistics (Booker et al 2020, Lotterhos, 2019 among others). The present study is nevertheless welcomed and timely, adding up on those previous by formerly showing how localPCA behave in simulated data with and without selection.
Reply:

Thank you for the comment. In Introduction, we refer to studies in which outliers of genome scan of genetic variation are interpreted mostly as the effect of selection. As you point out, many empirical studies have found an association between outliers and reduced recombination rate, and many (but not all) interpret and discuss this as linked selection. In Revision 1, we have tried to make this clearer (revision 1 P4 L53-64).

3- The relationship between low-recombination and PCA outilers is not fully explored. Most (All ?) PCA outliers, as defined by showing consistent patterns of variation in a MDS, are also regions of low

recombination including inversions. On the contrary, are there regions of low recombination that are not detected as outliers of PCA ? If yes, what are the genomic features that differentiate them from the ones being outliers ? Knowing that in both empirical and simulated data could help understanding, on the one hand, the power of this analysis, and on the other hand, what are the necessary conditions and what are additional factors possibly captured by local PCa analyses.

Reply:

Thank you for the comment. We agree that it should be both biologically and technically important to address low-recombining regions that are not outliers if they exist. The answers differ between species-wide and population-specific low-recombining regions.

Species-wide low-recombining regions.

First, in our newly added coalescent simulations (more details described below in response to the 5-th point and to Reviewer 2) and in Revision 1 P11 L166-177, Revision 1 Sup. Figs 19-23, Revision 1 Sup. Table 8), we observed that outlier regions were always detected at species-wide low/non-recombining regions and low-recombining regions were always outliers, suggesting a strong one-to-one relationship between reduced recombination rate and distinct patterns of genetic variation. This is consistent in our forward simulations of species-wide low-recombining regions: the distinct patterns representing haplotype structure persist until the population structure starts to emerge (Original and revision 1Fig. 4 BC, original Sup. Fig. 17 (revision 1 Sup. Fig. 24)). Consistently, in our empirical data of blackcaps, many outlier windows were those with the lowest recombination rates in a chromosome (Chromosomes 1 and 7 in Fig. 1 below). We note here, however, that we defined low-recombining regions applying a percentile threshold (20 percentile) per chromosome. This mild threshold was chosen because of the huge variation in recombination landscape among chromosomes and to make sure that population-specific reduction in recombination rate, which may not result in the lowest recombining region in a chromosome. This procedure resulted in many non-outlier windows which were labelled as “low-recombining” (Chromosome 21 in Fig. 1 below). These windows mostly reflect the relaxed threshold. In addition to this technical effect, we speculate that heterogeneous mutation rate along chromosomes may cause low-recombining non-outlier regions. At low-recombining regions, if the mutation rate is even lower, the window (defined by the number of mutations in) should contain multiple genealogies and thus exhibit local structure consistent with the population structure. This idea could be naively tested e.g. by associating the lostruct results in low-recombining regions with the physical window size in bp. However, such effects are difficult to discern from background selection (Charlesworth & Jensen, 2023), and as heterogeneous mutation rate is beyond our scope we decided against any further inspection.

Population-specific low-recombining regions.

In our forward simulations of population-specific low-recombining regions, we observed temporal transiency in distinct patterns of genetic variation. As we discuss in the original Sup. Fig. 20 (revision 1 Sup. Fig. 27), the primary axes of PCA within population-specific low-recombining regions represent cryptic haplotype structure, which requires the presence of linked mutations representing the haplotype. In other words, mutational noise (defined in original P15 L227-231, revision 1 P18 L289-294) plays a significant role in recoverability of cryptic haplotype structure at population-specific low-recombining regions. In line with the transiency observed in the simulations, our empirical results show a number of windows in which recombination rate is reduced in one population without showing distinct patterns in local PCA (chromosome 21 in Fig. 1 below).

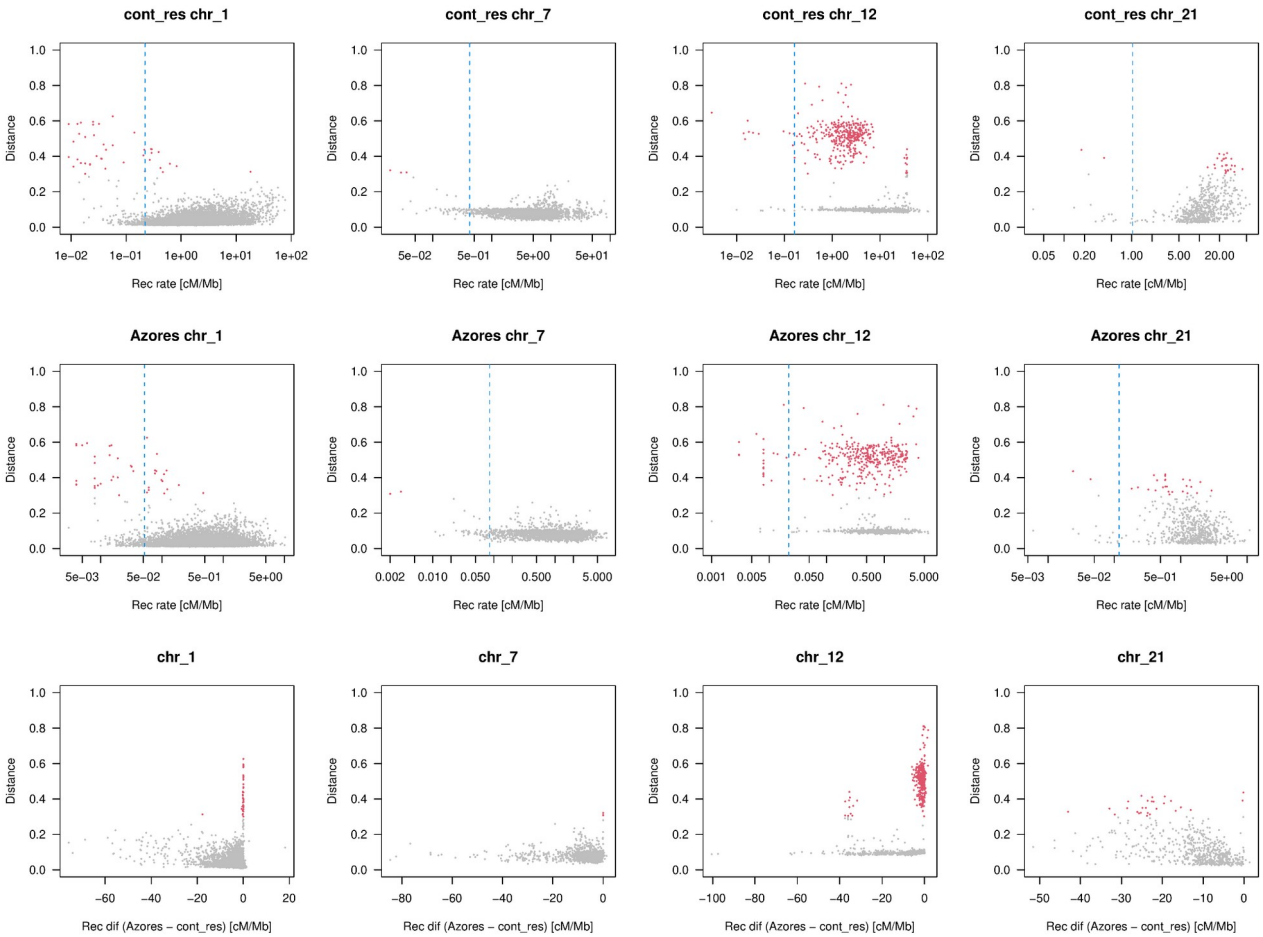

Figure 1. Relation between recombination rate and MDS values of lostruct. Four columns correspond to four exemplified chromosomes. Chromosome 1 exemplifies a chromosome with outliers overlapping species-wide low-recombining regions only. Chromosome 7 exemplifies a chromosome without outlier regions. Chromosome 12 exemplifies a large putative inversion. Chromosome 21 exemplifies a chromosome with only outliers overlapping population-specific low-recombining regions. The Y axis in all panels shows the longest distance along the 20 MDS axes in the haplotype-based analysis between the focal 1,000-SNP window (as defined in lostruct) and the mode of the distribution per chromosome. The X axis in panels of the first and second rows shows recombination rate of 1,000-SNP windows in cont_res and Azores populations. Red dots show outlier windows (Revision 1 P29 L568-571). The blue dotted vertical line shows the 20 percentile of recombination rate per chromosome per population windows below which were defined as low-recombining. The X axis in panels of the third row shows the difference in the recombination rate between cont_res and Azores.

Charlesworth, B., & Jensen, J. D. (2023). Population Genetic Considerations Regarding Evidence for Biased Mutation Rates in Arabidopsis thaliana. Molecular Biology and Evolution, 40(2), msac275. https://doi.org/10.1093/molbev/msac275

4- There is some inconsistencies in the text. Most of the results and the text explain that linked selection may or may not be present, which is fine and cautious, but some sections take shortcuts and claim the absence of selection. This may be misleading. I suggest sticking to the former rather than the latter. Additional tests than localPCA can and should be done to study selection – in the same way that additional proofs are needed to confirm a putative inversions detected by local PCA.

Reply:
Thank you for the comment. We edited Discussion accordingly, making possible linked selection clearer. Selection tests in the empirical blackcap data were in original PP13-14 L177-185, Sup. Figs 21-

23, Sup. Tables 8, 9, (revision 1 PP15-16 L223-231, Sup. Figs 28-30, Sup. Tables 9, 10). We now also include population genetic summary statistics (pi, Taima's D, and FST) for the data simulated under neutral coalescent with reduced recombination rate and different demographic histories. Models are described in revision 1 Sup. Figs 21-23). The results show increased variance in low-recombining regions.

5- There is no mention on the impact of population size, despite a choice for a low Ne in the simulation – 1000 individuals split in 3 populations). What could be the impact of such parameter? How could that explain the pattern in Island populations? Intuitively a lower Ne means less opportunity for recombination, less different genealogies…

Reply:

Thank you very much for this comment. We have now expanded our simulations to include variation in Ne after population splits, and also discuss this issue more elaborately in general.

In our original forward simulation study, we addressed how (evolution of) the recombination landscape affects local genetic variation. The purpose of the simulation was to evaluate the effect of the recombination landscape on the local genetic variation in general, instead of reproducing the parameters in the blackcap. We used forward simulation with SLiM instead of coalescent simulations such as msprime because population-specific recombination maps cannot be implemented in the latter. Although species-wide recombination reduction can be simulated with msprime, we had kept (in the original version) the simulators consistent between species-wide and population-specific reduction in recombination rate. The particular choice of population size in our simulation was to keep the simulation computationally manageable, and we scaled mutation and recombination rates and time accordingly, whereby we kept the simulation as simple and tractable as possible while keeping it biologically meaningful (original P32 L676-680).

Regarding the effect of demography, we now added neutral coalescent simulations using msprime under a series of scenarios differing in (species-wide) recombination maps, population structure, and demographic history (Revision 1 P11 L166-171, PP38-39 L823-844, Sup. Table 8, Sup. Figs. 19-23). In these simulations, the order of effective population size, the time of population split, mutation rate, recombination rate, and unbalanced sample size are tuned to our blackcap dataset. The results show that reduced recombination rate, but not demography (including reduction in Ne in some populations, as both reviewers suggested), causes outlier regions. Furthermore, the distinct pattern of genetic variation at low-recombining regions summarised with PCA consistently represent combinations of distinct haplotypes in both coalescent simulations with msprime and forward simulations with SLiM. Therefore, the distinct patterns at low-recombining regions simulated with SLiM is not due to scaling of parameters.

· Minor comment

Title : It feels slightly unclear and expected- any region will reflect a structure of haplotypes, but the haplotype length depends on recombination rate –

è Maybe « haplotype structure rather than population structure » to be more explicit ?

è Also given that the focus is on local PCA rather than other ways to study genetic variation, perhaps that can be explicit ?

.è Or highlight the importance of recombination rate rather than just low-recombining regions ?

Reply:

Thank you for the suggestions. Our preference would be to stick to what we have in our original submission. First, although the haplotype length is likely affected, we did not address this aspect in this manuscript thoroughly but focused more on the patterns of genetic variation. For the second suggestion, “Distinct patterns” implicitly contains the message of “instead of population structure”, so we think it would make the title redundant. Regarding the third suggestion, our emphasis is on low-recombining regions instead of general recombination landscapes including high-recombining regions, because in the limit of high recombination rate (i.e. all SNPs are under linkage equilibrium) there would not be outlier regions under neutrality.

L6 what does « too few genealogies » means ? too few generation to recombine ? too few ancestors ? too few distinct lines ? More generally, reading the abstract, the word genealogy may need a definition

Reply:

We apologise for the confusion. In the revision 1, “underlying genealogies” (original P1 L5) now reads “underlying genealogies representing local genetic ancestries” (revision 1 P1 L5-6).

L7 what does « distinct patterns of genetic variation » means ? perhaps « as displayed on PCA » ?

Thank you for your comment. With distinct patterns of genetic variation we generally refer to patterns “distinct from general population structure”. We specify this in our original P1 L1-2 as follows: “Genetic variation of the entire genome represents population structure, yet individual loci can show distinct patterns.” With this sentence we introduce that by “distinct patterns of genetic variation” we mean “patterns different from the population structure”. PCA is one of many ways to summarise the pattern of genetic variation. In our study, we use a PCA-based method because this approach very nicely fits our purpose (i.e. an explorative method to summarise the variation without predefined population labels), but we are interested in distinct patterns of genetic variation in general instead of a particular method of summarisation.

L10 « with reduced recombination rate » ? or rather with the recombination landscape » ?

Reply:
Thank you for pointing this out, we agree that our initial phrasing was unclear and we have edited the sentence as suggested, it now reads “Here, we associate distinct patterns of

local genetic variation with reduced recombination rates in a songbird….” (revision 1 P1 L9-11).

L36 « a sufficient number of variable sites » -> a sufficient number of *unlinked* sites. Best practices often recommend LD pruning before structure or clustering analyses.

Reply:

Thank you for this comment. The revised sentence now reads “Inference of population structure as well as other genome-wide analyses based on genetic variation take advantage of a sufficient number of unlinked variable sites …” (revision 1 P3 L35-37).

L42. Summarising with measure on the entire genome may not even be enough. For exemple, some regions of low-recombination (e.g. inversions) can not only take over local patterns but also global patterns (affecting a PCA on all the genome). Of course, I have our study on the seaweed flies in mind (Mérot et al 2021) but the same has been observed in many species, particularly marine ones (cod,

capelin, etc). This is one of the reasons why it may also be useful to explore local patterns of PCA and how the heterogeneity in recombination rate impacts the structure of genetic variation.

Reply:

Thank you for pointing this out. We are aware that some large low-recombining regions (under selection) affect the population structure based on the whole-genome variants. However, the key message of this particular paragraph is that the random fluctuation of genealogies can be (and has been) removed by using many unlinked loci, and we are worried that adding these special cases here would make the paragraph confusing. We therefore decided to include this information in revision 1 P4 L59-61, which reads “Distinct patterns at low-recombining regions can influence the chromosome-wide (Knief et al., 2016; Neafsey et al., 2010) and even genome-wide population structure (Mérot et al., 2021).”, within the paragraph focusing on distinct patterns.

L56-58 : « Distinct patterns of local genetic variation identified with genome scans are often attributed to the effects of selective factors instead of randomness (Burri, 2017; Mérot et al., 2021) based on the assumption that the genomic intervals are large enough to eliminate random fluctuation of genealogies (Li & Ralph, 2019) ». Here the references given do not support this assertation. On the contrary, the method of local PCA from Li & Ralph 2019, and used in Mérot et al 2021 precisely does the opposite. It relies on distinct patterns in PCA variation to uncover genomic regions with underlying haplotype structure. This haplotype structure may be due to several factors including low recombination (centromeres, chromosomal inversions, heterogeneity in recombination landscape) with or without linked selection – like in the blackcap system in fact! For exemple no selection is needed to explain the haplotype structure due to inversions (and thus the specific local PCA), simply the reduction of recombination is enough. Then the inversion may or may not be under selection. The value of the present paper is to explicitly simulate cases with and without linked selection to explore how an analysis of localPCA behaves.

Reply:
We apologise for the confusion. We did not cite Li & Ralph 2019 there to refer to local PCA but instead for specific sentences in their introduction: “*These realized patterns of genetic relatedness summarize the shapes of the genealogical trees at each location along the genome. Since these trees vary along the genome, so does relatedness, but averaging over sufficiently many trees we hope to get a stable estimate that does not depend much on the genetic markers chosen.*” (Li & Ralph 2019). This assumption is made implicit in most population genomics papers and we could not cite other papers specifically on this. We would be happy to be informed of such papers/books.
We now carefully chose papers interpreting outlier regions in favour of linked selection with only minor consideration of the neutral effect of low-recombining regions (revision 1 P4 L61-64).

Results : not a single words about the genome assembly ? About the extensive confirmation of inversion breakpoints ?

Reply:

Thank you for the comment. We now included a paragraph for genome assembly (revision 1 P6 L96-109).

L98 : windows of 1000 SNPs (worth mentioning here because methods are at the end)

Reply:

Thank you for the comment. We restructured the paragraph with more information of the procedure, and included the window size as suggested (Revision 1 P7 L126-127).

L99 : Outliers in the nMDS analysis are not exactly windows with distinct patterns, they are rather groups of windows with the same exact pattern which differs relatively to a background of windows with heterogeneous patterns)

Reply:

Thank you for this comment, we have rephrased this part to be clearer. Now the corresponding sentences read “Briefly, lostruct performs PCA in sliding genomic windows, and dissimilarity of PCA among windows are summarised with multidimensionality scaling (MDS). Distinct patterns of genetic variation of windows relative to the background are represented by extreme values along the MDS axes. Multiple windows with correlated patterns of genetic variation distinct from the population structure are represented by extreme values along the same MDS axis.” (P6 L119-122).

L101 : which threshold ? – give briefly parameters

Reply:

We appreciate you pointing this out and we understand this calls for more explanation and asks for a brief reference to parameters. Now the corresponding sentences read “We performed `lostruct` on both genotype and phased haplotype data with window size of 1,000 SNPs. We identified outlier windows by applying threshold MDS values (the mode of the distribution ± 0.3). We further identified genomic regions with distinct patterns of genetic variation by finding genomic intervals longer than 100 kb with at least five outlier windows based on the same MDS axis and merging the intervals based on the genotype- and phased haplotype-based approaches.” (Revision 1 P7 L126-131)

L302 : 32 regions from Xmb to Xmb (mean length) each including X SNps to X SNPs

Reply:

Thank you for highlighting this. We have rephrased this part and the corresponding sentences now read: “This yielded 32 genomic regions with distinct patterns of variation (hereafter “outlier regions”, Fig. 2D, Sup. Table 3, Sup. Fig. 2). Their size ranged from 0.12 to 8.11 Mb (mean and median of 0.71 and 0.29 Mb), and each region contained 5,000 to 356,000 SNPs.” (Revision 1 P7 L131-134).

L104 : low-recombining regions, defined as regions with a recombination rate below X ?, were ….

Reply:

Thank you for the comment. The corresponding sentence now reads “Comparing the genomic distribution of these outlier regions to population-level recombination maps, we found that low-recombining regions (nominally recombination rate lower than the 20 percentile of the the recombination map for each chromosome) were significantly enriched in the outlier regions (permutation tests, p-value < 0.001 (Sup. Fig. 10)).” (Revision 1 P7 134-137).

L106 : how many outliers regions coinciding with species wide vs. Pop specific low rec regions ? Are there outliers regions outside low-rec regions or not a single one ?
Reply:

Thank you for this question. Inspired by your and Reviewer 2's comments, we improved our pipeline to define low-recombining regions and to label outliers according to overlaps with low-recombining regions (Detailed below and revision 1 PP31-32 L614-651). Nine outliers overlap with species-wide low-recombining regions, 11 outliers overlap with population-specific low-recombining regions, 2 outliers have no overlaps with low-recombining regions.

L119-123 : What are the arguments/results supporting inversions beyond LD ? What are the patterns of LD consistent with non-inversion haplotype blocks ?

That was very interesting to fully see the exploration in supplementary materials. In Fig S10 I am not sure one can neither exclude or support an inversion-like region for outliers 6,14 and 28. In particular, the low frequency of the B allele at outlier 6 and outlier 14 probably means that most SNPs are within AA rather than between A and B, possibly explaining the persistence of LD in AA. Those regions are also very small 100-300 kb, with additionnal geographic structure, making it difficult to pinpoint the cause of the three clusters. I agree that Outlier 30 is much bigger (1.5Mb) and more typical of simple polymorphic inversion. [I realised that this comment is useless for a review, please simply try to be more explicit for the readers about what are the reasons that suggest inversion-like recombination reduction vs. Low-recombination due to other possible mechanisms]

Reply:

Thank you for your feedback. The clusters of individuals simply represent combinations of distinct haplotypes (which we refer to as: " due to presence of two distinct segregating haplotypes" in Revision 1 P9 L157). Simulation of a species-wide low-recombining region is in support of this interpretation (" The low-recombining regions exhibited three, six, or more clusters of individuals resembling our empirical results. The clusters of individuals represented genotypes consisting of different combinations of ancestral haplotypes (Sup. Fig. 25)." (Revision 1 P11 L186-188), but we did not explicitly mention this to avoid complication in the text.

L126 « spread in PCA projections » -> on PC1 and PC2. Does this pattern holds true over more PCs ? Variance is hardly interpretable with a subset of PCs. The authors may be interested in checking Elhaik preprint. The title and message are very extreme but there may still be a few things to take from it regarding the interpretation of PCA.

Why most Principal Component Analyses (PCA) in population genetic studies are wrong. Eran Elhaik bioRxiv 2021.04.11.439381; doi: https://doi.org/10.1101/2021.04.11.439381

Reply:

Thank you for this comment. Yes, we checked other axes than the first two PCs in the outlier regions with population-specific reduction in recombination rates. This pattern (based on population-based colouring on PCA) held up to PC2 or PC3 at outlier regions with population-specific low-recombination. In theory, the same pattern should not hold for multiple PC axes because they are orthogonal to each other. Our results for the population-specific low-recombining regions indicate that the first few PC axes capture distinct variance among different cryptic haplotypes within the same populations. As mentioned in Discussion, these axes represent distinct ancestral haplotypes contributing to current haplotypic variation. The fact that the spread of individuals of the low-recombining populations occurs at the first few PC axes indicates that they represent a few, but not many, cryptic haplotypes.

Figure 4 : I suggest writing on the side B/C low rec ; D high rec, for quick reading. Or adding a little drawing of recombination landscape and pointing where the windows looked at in B/C and D are.

Reply:

Thank you for the suggestion. We included indicators of “low-rec.” and “normal rec.” beside the panels in Figures 4 and 5.

L204 « distinct patterns of genetic variation » - as observed through PCA ? Those outlier regions are may or may not distinct for other ways of evaluating genetic variation (pi, Fst, heterozygosity, etc etc)
Reply:
Thank you, this question is essentially whether the distinct PCA at low-recombining regions are due to general effect on genetic variation or something PCA is especially susceptible to. Hudson (1983) shows that the mean and variance of the number of polymorphic sites between two sequences (E[S] and Var[S], respectively) under the two-locus model of (neutral) coalescent with recombination are

$$E[S]=\theta$$

$$Var[S]=\theta+\theta^2\frac{2}{\rho^2}\int_0^{\rho}(\rho-x)f_2(x)\,dx$$

where θ is the population mutation rate, ρ is the population recombination rate, and $f_2(x)$ is the correlation of the total branch length of the genealogy between the two loci expressed as

$$f_2(x)=\frac{18+x}{18+13x+x^2}.$$

E[S] is not affected by recombination rate, while Var[S] under coalescent with recombination approaches that of standard coalescent (without recombination) $\theta + \theta^2$ as ρ approaches 0, and it decreases monotonically to θ as ρ increases, which is in line with increased variance in polymorphism at low-recombining regions. To our best knowledge, no simple expression of variance of other classical summary statistics have been derived. Correlation between (reduced) recombination rate and (extreme) summary statistics in empirical data is difficult to interpret due to the potential (and likely) effect of background selection. However, our neutral simulations show that variance of summary statistics is elevated at low-recombining regions (Revision 1 Sup. Figs 21-23), in line with Hudson (1983). These lines of evidence indicate that outliers based on local PCA reflect the general effect of low recombination rates on (the variance of) genetic variation, rather than something specific to PCA.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology, 23(2), 183–201. https://doi.org/10.1016/0040-5809(83)90013-8

L209 « instead of selection » inaccurately reflects the last paragraph. Apparently the contribution of linked selection is not necessary to make this structure but can be present but should hasten the separation of populations rather than haplotypes (If I understood correctly Fig S24). Perhaps « reflects primarily reduced local recombination rates ? » (without a necessary contribution of selection)

Reply:
Thank you for pointing out this is a little confusing in its original phrasing. The corresponding sentences now read “ We showed empirically that genomic regions with distinct patterns of genetic variation identified by a population genomic scan based on principal component analysis (PCA) highly overlap with low-recombining genomic regions (Fig. 2). With simulations, we showed that although selection may affect the amount and pattern of local genetic variation around the target locus, the distinct patterns of genetic variation represented by PCA at low-recombining regions can be primarily explained by haplotype structure due to reduced recombination rate (Figs. 4, 5).” (revision 1 P17 L272-278).

L271-295 : This matter of variance in regions of low recombination and the impact on genetic statistics (particularly when summarized by windows) has interestingly been highlighted by Booker et al. Perhaps a good reference for this idea.

Booker, TR, Yeaman, S, Whitlock, MC. Variation in recombination rate affects detection of outliers in genome scans under neutrality. Mol Ecol. 2020; 29: 4274–4279. https://doi.org/10.1111/mec.15501

Stevison, LS, McGaugh, SE. It's time to stop sweeping recombination rate under the genome scan rug. Mol Ecol. 2020; 29: 4249–4253. https://doi.org/10.1111/mec.15690

Reply:

Thank you for pointing us to the highly relevant paper. We now refer to it in our manuscript.

L305-311 Since those are islands population, may they have a lower effective population size than elsewhere ((or a past bottleneck) ? How does low Ne amplifies the haplotypic structure ?

Reply:

Thank you for the comment. In the newly added coalescent simulations (described above, response to Reviewer 2, and Revision 1 P11 L166-177, Revision 1 Sup. Figs 19-23, Revision 1 Sup. Table 8), we included scenarios in which some populations experience a reduction in effective population size. Outliers were detected always and only when recombination rate is reduced, irrespective of the presence of population structure or demographic history.

L364 Simulations from Lotterhos 2019 show precisely which statistics are affected by low recombination and which ones are not. In particular everything affected by LD (PCA, clustering, window-average) are particularly sensitive. Perhaps more nuance is needed here to recall that analyses remain possible !
Reply:

Thank you for the comment. We have included a brief description of what could happen at low-recombining regions purely by chance reflecting high variance and what less likely happens without selection. Revision 1 P24 L433-436 now reads "For instance, apparent outliers in only few (pairs of) populations at a low-recombining region may reflect high variance, while high variance at low-recombining regions alone cannot explain signals occurring in many (quasi-) independent populations or species at a low-recombining region."

L443-L444 : No batch effect between the 69 and 110 blackcaps sequenced separately ?

Reply:

Thank you for the comment. Both the published data of 110 individuals and new data of 69 individuals ("sets" hereafter) were sequenced in multiple batches/in separate runs and on separate lanes ("true batch" hereafter) of Illumina sequencing (a total of ten true batches with three different platforms in two sequencing facilities). In this study, they were mapped to the same reference and SNPs were called jointly, meaning there are no effects by calling SNPs separately. Within each true batch, we aimed to include individuals from different populations to avoid nesting of the true batch in populations and sampling locations as much as possible. In the dataset used in this set of analyses you correctly point out there might be a batch effect between newly added and previously used samples. This could be of concern as new sampling locations were added in the new set, and some populations/locations are indeed unique to the newly added set. This means that for some true batches, these are nested in populations/locations. In Figure 2 below, we colour-coded individuals from the published and newly

added data on the PCA of original Fig. 2A. The two sets within the Canary and Madeira populations are weakly separated along PC1 (Figure 2a below), indicating the possible presence of a weak “batch effect” as highlighted in the comment by Reviewer 1. Stronger separation within the Cape Verde population along the same PC1 pointed out by Reviewer 2 (Figure 2b below) separates the two sets to a greater extent. However the individuals added to the new set were sampled on different islands of Cape Verde (the individuals in Delmore (2020) were sampled on Santiago (n = 4) and Fogo (n = 1), and the new individuals were sampled on Brava (n = 1), Santiago (n = 1), Sao Nicolau (n = 2), and San Antao (n = 2)). This was done on purpose in order to allow for better resolution within the archipelago as island colonisation of this species is separately analysed in another paper in preparation. A naive linear modelling of genetic variance along PC1 (the product of the eigen pair) with population, location, and set (formula of var_pc1 ~ population : location + set) shows insignificant effect of the sets (slope = 0.0032, p-value = 0.33). These results indicate that the batch effect by the sets may be present in PC1 which weakly separates individuals from the same location of the same population in different sets, but the “batch effect” by sets is not significant at the level of the whole PCA and should thus not compromise the analysis of this manuscript. In addition, this effect should be distributed uniformly throughout the genome, and thus it should not affect our analysis based on local PCA. Any biological interpretation of evolutionary history based on PCA should indeed be carried out with caution, but as mentioned, this is done in a separate paper, which is currently under preparation.

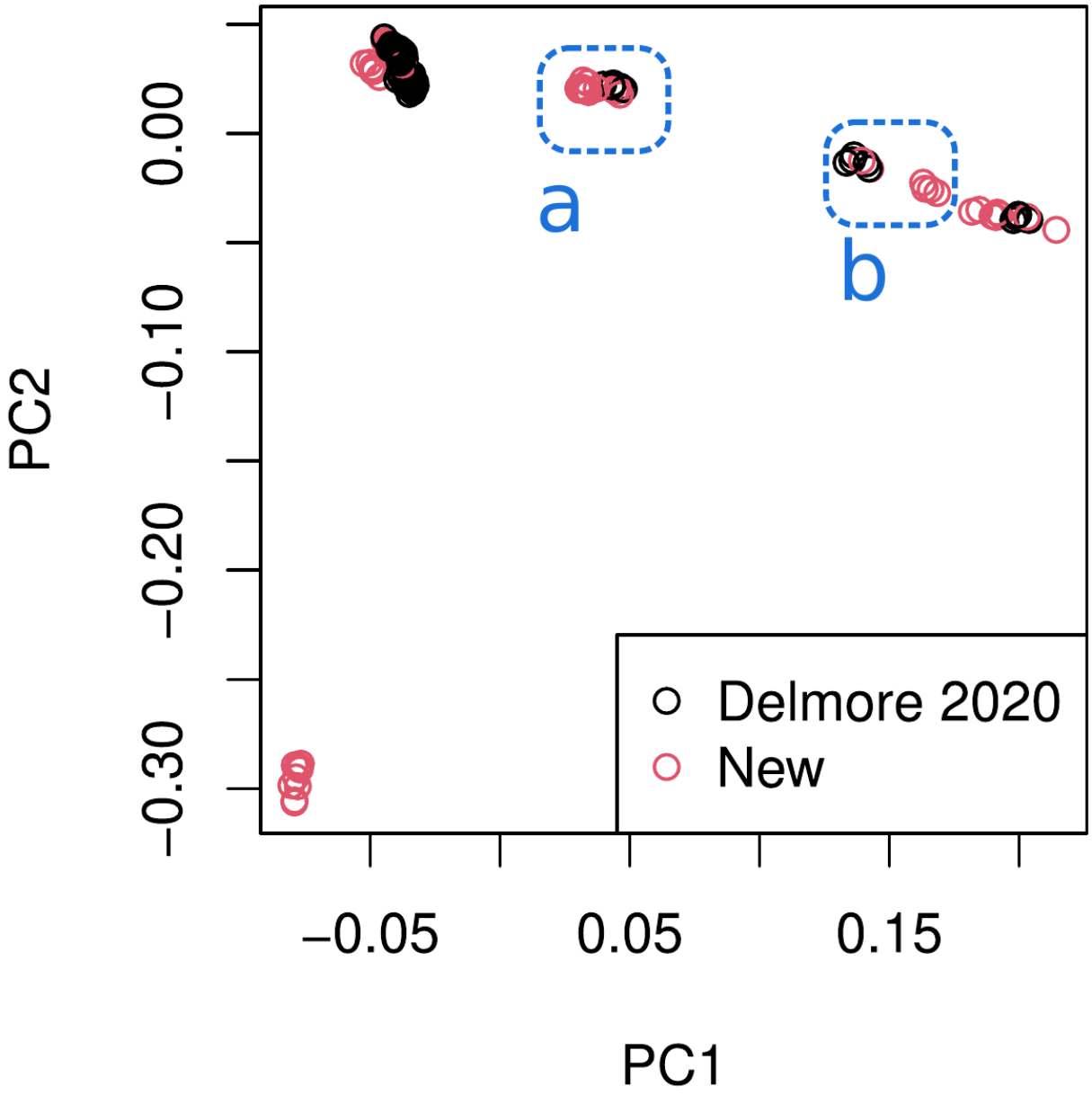

Figure 2. Batch effect on PCA. Original Fig. 2A was replotted with a colour-code by whether they were in the published set. a. Canary and Madeira. b. Cape Verde.

L451 : what is the depth of sequencing ? the realized coverage ?

Reply:

Thank you for the question. The information is now included in Sup. Table 1 and briefly mentioned in revision 1 P28 L535-536, which reads “The minimum and median depth were 7.8X and 20.1X, and the minimum and median coverage were 0.88 and 0.97.”

L5514 : confirmation of inversion and breakpoints : why is that not reflected in the results ?

Thank you for the question. We decided that the most relevant information regarding the putative inversions in this manuscript is that the recombination rate is reduced and it is segregated in many populations, and therefore we transferred all attempts to identify breakpoints to the supplementary

materials as we felt this would help to keep the main focus of the manuscript more concise, and we were worried it might veer the manuscript off-topic. However, given the referee comments on this specifically, we would be happy to instead integrate and present these results in the main results section, if the referees feel they are essential to the manuscript.

Review by anonymous reviewer 1, 19 Dec 2023 15:59

Ishigohoka et al., examine regions with distinct patterns of genetic variation in the blackcap genome and show that these often correspond to parts of the genome with a low recombination rate. They examine this property with simulations, and discuss the implications of these regions. Overall, this is an interesting paper exploring a seemingly simple but often overlooked concept. It adds to the recent body of literature stressing the importance of considering local recombination rate and the impact it can have on certain common measures.

I do think the paper could go a bit further, as although it shows that these regions exist, the implications of them are still just suggested rather than demonstrated. I have some suggestions for additional analyses that would strengthen the conclusions plus some important clarifications on the methodology. I also found some parts hard to follow and have a few ideas for improving clarity.

The abstract is a nice succinct overview of the paper. The introduction is also well-written and explains concepts well.

The supplementary figures are out of order with when they are mentioned in the text, and some are only mentioned in the methods section. This made it quite confusing to follow some of the analyses.

Line 11: Include Latin name of blackcap here?

Reply:

Thank you for the suggestion. We have added this information and the corresponding sentence has been edited accordingly.

Figure 1- A nice Figure, explaining the concept simply. Clarify that (D) is a PCA in the legend.
Reply:

Thank you for the kind words, especially because this figure underwent multiple rounds of revisions prior to submission. To improve comprehensiveness, we have now added the following sentence to the legend. "The realised genetic variation can be summarised and visualised with various methods such as PCA (D)". We avoided specifying PCA within the figure (e.g. writing “PC1” and “PC2”) because conceptually this can be other methods of summarisation.

Line 41- ‘Usually’ projected onto a few major axes- some analyses use many more axes.
Reply:

Thank you for the suggestion. The corresponding sentence has been edited accordingly.

Line 56- Would be nice to have a few more cited examples here.

Reply:

Thank you for the comment. We restructured the paragraphs and included some more references (Revision 1 P4 L47-70).

Line 81- confusing sentence- remove one of the ‘genetic variation’s

Reply:

Thank you for pointing out this is confusing, the corresponding sentence now reads "We further investigate the patterns of genetic variation in outlier regions and associate them with the prevalence of recombination suppression across populations."

Line 94- I couldn’t find how the whole genome PCA was run in Methods - e.g. which software did you use, did you filter for LD and how? Overall, I often find that smartPCA from eigensoft is better than PLINK and could be worth a try here (although it might not make too much of a difference).

Reply:

Thank you for pointing this out and for your suggestion. We are aware of different implementations of PCA other than what we used, such as smartPCA. The most consistent approach in our study would be to use the same implementation throughout the study between local PCA and any other PCA (i.e. sticking to the function in lostruct). Although it is possible, we did not use the function for PCA in lostruct (lostruct::pca_cov, which applies the base::eigen function on the covariance matrix computed by the base::cov function) for summarising the structure in the outlier regions and whole-genome with many SNPs because computation of the covariance matrix is time- and memory-consuming in large genomic regions. Instead, we used PLINK for summarising genetic structure (genome-wide and outlier regions consisting of multiple windows) for ease of analysis. Comparison of the performance between smartPCA and what we used is not straightforward to control for the conditions: smartPCA has an algorithm to detect and remove outlier individuals iteratively to better capture the population structure, which is not included in lostruct. Therefore, using smartPCA on outlier regions might not be a good representation of the distinct structure detected by lostruct. Nevertheless, comparison of PCA implementations between lostruct and PLINK is still valuable to make sure of the consistency between outlier detection and summarisation (described below).

Thank you for pointing out that the description of whole-genome PCA was missing. The procedure of whole-genome PCA is now included in Materials & Methods (revision 1 P28 L548-550). As no explicit requirement of SNP filtering is described in lostruct, we naively applied lostruct on quality-filtered and phased (and imputed) VCF without LD-pruning. To summarise the pattern of local genetic variation at outlier regions defined with lostruct, we used PLINK. We did not perform LD-pruning here either to keep the set of SNPs same with those used in lostruct. Although this naive use of SNPs without LD-pruning appears to be often the case in many publications applying lostruct, we agree with the comment that LD-pruning could be fundamental and the effect of the lack of LD-pruning should be validated.

To evaluate the above two points (consistency between lostruct and PLINK, and effect of the lack of LD-pruning), we compared 1. PCA with PLINK using all SNPs of all autosomes (original Fig. 2B), 2. PCA with PLINK using LD-pruned SNPs (using PLINK with “--indep-pairwise 1000 1000 0.2”) of all autosomes, and 3. PCA with lostruct (cov_pca function) using all SNPs of all autosomes. 1 vs 2 addresses LD pruning, and 1 vs 3 addresses PLINK vs lostruct. All the three were consistent with each other (Figure 3 below). This indicates 1. summarisation with PLINK reflects what is detected by lostrust,

and 2. the lack of LD pruning has little effect in our study.

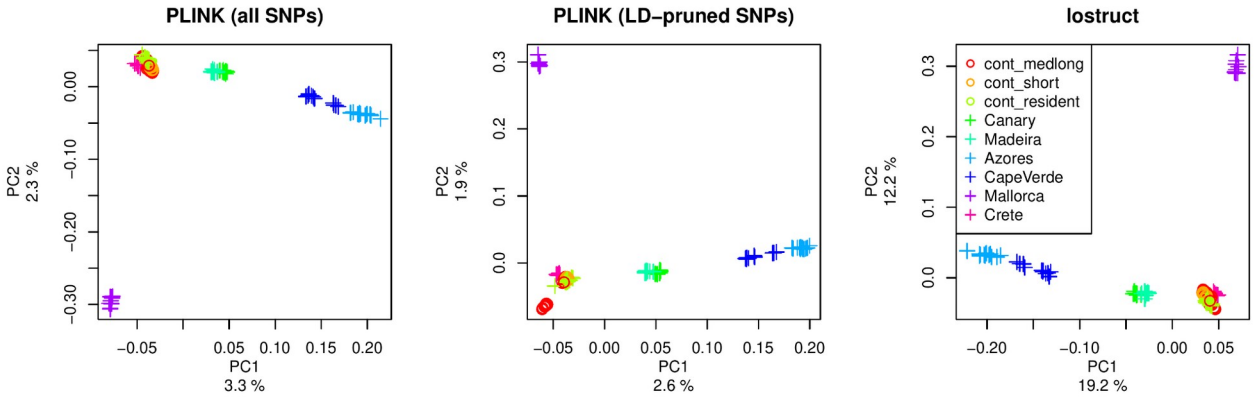

Figure 3. Whole-genome PCA of the blackcap data comparing the effects of implementations (PLINK vs lostruct) and LD pruning. The difference in the variance explained by PCs between PLINK and lostruct is likely due to the number of PCs specified (20 for PLINK and 2 for lostruct. lostruct did not run for the number of PC axes greater than 2 for the very large variance-covariance matrix).

Line 98- Include size of window here. Also line 102- include brief definition of how an outlier was classified.

Reply:

Thank you for the suggestion. We included a brief description of the procedure including the window size and outlier calling in Results (Revision 1 P7 L126-131).

Figure 2A- It looks like there are 2 different Cape Verde populations from the PCA- have you tried splitting these and seeing if there are any population specific low-recombining regions within Cape Verde? Also, for ease of visualisation, 2A could be larger and have size of point correspond to number of individuals.
Reply:

Thank you for these comments. The Cape Verde population indeed shows substructure in genome-wide PCA. This corresponds to two sets of islands within the archipelago (also described in response to Reviewer 1). Analysis on more detailed population history is carried out in another paper under preparation. It would be certainly interesting to investigate recombination map divergence between these closely related subpopulations. However, currently we are limited with the number of samples for each island of the archipelago to perform such comparison.

Thank you for the suggestion on Figure 2A. We decided to leave the size of Fig 2A the same to keep the other panels for the main results as large as possible. We included the sample size information in Fig 2A accordingly, where the point size reflects the square root of the sample size (i.e. the area of the circle reflects the sample size).

Figure 2D- I am confused by how many outlying regions overlap with low-recombining ones- do they all? This should be clarified more clearly in the text

Reply:

We apologise for the confusion. In the original manuscript, we assigned species-wide or population-specific low-recombining regions to the outlier regions by visual inspection of the plotted recombination maps and PCA after confirming significant association between outlier regions and low-recombining regions with a permutation test per population. We appreciate your comments here and below (regarding the definition of overlaps in the permutation test), and improved our pipeline.

This procedure (described below and in revision 1 PP31-32 L614-651) resulted in 19 outliers overlapping with species-wide low-recombining regions, 11 outliers overlapping with population-specific low-recombining regions, 2 outliers without overlaps with low-recombining regions. These numbers are now included also in Fig. 2D. These assignments contain a few changes from the original manuscript. We re-did all statistical tests and visualisation accordingly and updated the supplementary materials. The general conclusions were not affected.

Figure 2F- Label which population has the low recombination rate (can see it is Azores, but would help clarity)
Reply:

Thank you for this comment. In the original manuscript, Fig. 2E&F contained recombination maps of two populations (cont_medlong and Azores) only to keep the figures as simple as possible. The two outliers on chromosome 14 (original Fig 2F, green shades) overlap with population-specific low-recombining regions in Azores and Cape Verde, and labelling this information would make the panel busier and even more confusing (with one population missing (Cape Verde)). In Revision 1, we include both Azores and Cape Verde recombination maps, and specifically mention in the legend that the two outliers in green overlap Azores&Cape Verde-specific low-recombining regions.

Figure S2- Haplotype-based analyses are often more powerful than genotype-based- is there much difference in results when just using the haplotype version of lostruct?
Reply:

Thank you for the comment. lostruct does not specifically have halotype-based option, but we prepared input position-by-sample matrix based on phased haplotypes for haplptype-based analysis. As validated in Revision 1 Sup. Figs 3-4, the haplotype- and genotype-based results of lostruct were highly correlated, and outlier regions were mostly overlapping between the two. To make this clearer, we added in Sup. Fig 3 gray shades to show positions of final (merged) outlier regions to help compare between positions of outliers detected in haplotype- and genotype-based results.

Methods line 503- How many of these regions were discarded and what proportion? How was similarity to the whole genome PCA judged? Did you try any other thresholds to determine outliers and check how the proportion that looked ‘normal’ changed- e.g. do you need a more conservative threshold?
Related to line 115- ‘These clusters did not clearly separate populations’ surely this is because you removed all the ones that did?

Reply:
Thank you for the questions. We removed seven regions besides the final 32 outlier regions. This decision was based on visual comparison of PCA (colour-coded by populations) of each outlier region with whole-genome PCA. This process could be potentially automated by additional summarisation taking the population label into account, but this is beyond the scope of our current study. These removed regions are shorter (mean: 212,345 bp, standard deviation: 81,960.5 bp) than the final 32 outlier regions (mean: 712,196 bp, standard deviation: 1,399,758 bp). The critical threshold in our pipeline in respect to these presumably false positive outlier regions is the number of outlier windows to be regarded as an outlier region. We defined an outlier region a region with at least five outlier windows

along the same MDS axis to remove noise in a few windows. By increasing this value, we would reduce the number of false positive outlier regions, but we would also miss short outlier regions. On the contrary, a smaller value of this threshold would include small regions with distinct structure, but it would also increase false positives.

The distinct patterns representing the haplotype structure could still separate populations. For example, a PC axis could separate migrant (cont_medlong, cont_short) and resident (island populations and cont_resident) populations instead of representing separation between different island populations versus the continental populations as in the genome-wide PCA.

Figure S6- This is quite a busy Figure- maybe it would be clearer if the population-specific ones were in a separate figure? Also I'm unsure about some of the categories, e.g. what category is outlier_3_1? Also some of the 'mixed_individuals' PCAs seem quite similar in shape to the '6 loose clusters'.

Reply:

Thank you for the suggestion. The original organisation of this figure was to make it easier to refer to them from in the main text, and to appreciate the variation in the patterns within each class instead of strictly classifying them into different subclasses. Based on the suggestion and updated labels of each outlier, we split the figure into three for outliers overlapping species-wide low-recombining regions (Sup. Fig. 6), those overlapping only with population-specific low-recombining regions (Sup. Fig. 7), and those without overlaps with low-recombining regions (Sup. Fig. 8) without specifying further subcategories.

Figure 3A- scale and units for the LD is missing (and in the supplementary figures).
Reply:

Thank you very much for pointing this out. The colour scale is now added to both figures.

Methods line 540- Could do with some more explanation of the permutation test- e.g. how did you calculate overlap, did it have to cover a certain proportion of the length of the outlier region, or 100%, or just any overlap? E.g. Figure S7 chr 30 species-wide outlying region does not seem to have a low recombination rate in med_sw and cont_res and similarly for chromosome 28?

Reply:

Thank you for this comment. In the original manuscript, the overlaps were of any length. However, as described in response to your question on Figure 2D, we improved the pipeline and now we count the number of overlapping base pairs, instead of counting the number of intervals overlapping at least by 1 bp.

The outlier on chromosome 30 (outlier_30_1) is one of the putative inversions segregated in multiple populations. We kept them in the "species-wide low-recombining" regions, because heterozygotes, in which recombination is suppressed, are present in many populations. Recombination maps in original Sup. Fig. 7 (revision Sup. Fig. 9) were performed before investigation into the putative inversions without accounting for the inversion genotypes, and because multiple AA individuals were included, the inferred recombination rates within outlier_12_3 and outlier_30_1 are not as low as other species-wide low-recombining regions. Suppression of recombination at these loci between A and B is evident in the original Sup. Fig 12 (revision 1 Sup. Fig. 14). Your point on outlier_28_1 is absolutely correct and also supported by the new pipeline. Outlier_28_1 is now labelled "population-specific (Azores, Cape Verde)".

Methods section on Inversion breakpoints- I could not find this section mentioned in the main results section ever? Should be added as a paragraph into results section or removed from paper?

Reply:
Thank you for the suggestion. We decided that the most relevant information regarding the putative inversions in this manuscript is that the recombination rate is reduced, and therefore we transferred all attempts to identify breakpoints in the supplementary materials as we felt this would help to keep the main focus of the manuscript more concise, without veering the manuscript off-topic. However, given the comments on this specifically by both Reviewers, we would be happy to instead integrate and include these results in the main results section, if the Reviewers and the recommender feel they are regarded to be essential to the manuscript.

Line 135- mention which simulator used.

Reply:

Thank you for the comment. The corresponding sentence now reads "To address how species-wide and population-specific reduction in recombination rate affect the patterns of genetic variation over time, we performed forward simulations using SLiM (Haller & Messer, 2022)." (revision 1 P11 L178-180).

Line 144- I wouldn't say 'population structure emerged' completely, especially not when compared to 4D, maybe just some clustering by population? Also in Figure 4, what do the time points correspond to when compared to Figure S17? Also, I am a little confused about the difference between Figure S17 and the top row of Figure S24- why does the latter continue until t=1600, would the results in s17 look similar if the time was increased?

Reply:
Thank you for the questions. The corresponding sentence now reads "The distinct patterns representing haplotype structure persisted until population structure started to emerge along the PC axes (Fig 4B, C)." (revision 1 P11 L188-190). In Fig. 4, the time points are shown as N generations where N in our simulation is 1,000. We would be happy to change it into generations (i.e. t = 0, 50, 1000 [gen] in Fig B-D) as done in original Fig S17 if this is easier to interpret.
In the main figures, we generally tried to keep the number of columns up to three for better readability. Specifically, in Fig. 4 we picked three time points such that the first shows the initial states, the second represents a time point at which the pattern is different between low- and normally recombining regions, and the third represents a time point at which population structure is observable at both low- and normally recombining regions. The time points in original Sup. Fig. 17 were decided so that the variance among replicates can be visible. For example, at t = 600 [gen (=0.6N)], population structure has started to emerge in sim00 and sim08, while in other replicates haplotype structure is still primarily represented. In the original Sup. Fig. 24, the time points were decided to show the difference in the rate at which the population structure starts to emerge among different DFEs. The specific question on whether the original Sup. Fig. 17 would look at later time points (e.g. t = 1,600 [gen]) the same as in the original Sup. Fig. 24 is absolutely correct, because the case of DFE with no deleterious mutations represents the same simulations shown in the original Sup. Fig. 17. Although we are limited in space even in the supplementary figures to show hundreds of time points in 100 replicates, all scripts with seeds necessary to reproduce the data are accessible on Zenodo.

Line 152- Not sure about point of two different scenarios- clarify why they are being compared?
Reply:

Thank you for the comment. The corresponding sentences now read "Three populations (pop1, pop2, and pop3) and their ancestral population had 1,000 diploid individuals, and pop1 evolved a reduced local recombination rate. We considered two cases with respect to when the population-specific reduction in recombination rate is introduced: before or after differentiation of populations. In the first scenario (Sup. Fig. 26), recombination suppression was introduced at the same time as the three populations split, while in the second scenario (Fig. 4A) recombination suppression was introduced 4,000 generations after the split." (revision 1 P14 L198-204).

Figure S20BF- what is the Y axis?

Reply:

Thank you for this question. The Y axis corresponds to sequences/haplotypes. We now include this information in the figure (revision 1 Sup. Fig. 27).

Selection section:

Line 192- I could only find PCAs of selection on top of regions with a low recombination rate, none of the effect of selection on the normally recombining regions (e.g. in Figure S24)? Overall this whole section could do with more explanation, it is quite brief and I found it difficult to work out what you have actually shown.

Reply:

Thank you for this comment. The corresponding supplementary figure (revision 1 Sup. Fig. 31) now has a panel for the normally recombining chromosome. We also included brief descriptions of the simulations we performed with selection and reduced recombination rate (Revision 1 PP16-17 L236-262).

A key (and quite simple) analysis missing from this section would be to run the selection tests (Tajimas D, pi, Fst) on the simulations with a region of reduced recombination rate and no actual selection. If these measures are biased in the outlying regions it would be great evidence for your Discussion/Implication sections (e.g. line 364). I know you show that the low recombining regions in the blackcap genome are often under selection, but surely a big aim of the paper is to show that this could be biased by the regions themselves?

Reply:

Thank you for the suggestion. In the revision 1, we now include sliding window analyses of summary statistics (pi, Tajima's D, and FST) on neutrally simulated data with coalescent under different demography (Sup. Figs 21-23). The means of these summary statistics were affected by the demography (as they should) but not by reduced recombination rate, while the variance of these summary statistics was greater at low-recombining regions. This indicates that the observed patterns in the blackcap data cannot be explained without selection. Importantly, however, our simulations of reduced recombination rate and selection indicate that haplotype structure detected by local PCA primarily reflects reduced recombination rate.

Overall the Discussion needs more references to Figures that show each of the points mentioned (e.g. line 239, which figures show this), and also more citations (e.g. the first line needs citations of which studies it is mentioning). Also I found the first section of the discussion a little long, with the same concepts explained multiple times. The section on recombination landscape as a driver of evolution is interesting, and the implications are well explained.

Reply:
Thank you for the comments. We slightly restructured the first part of Discussion, included reference to main figures, and added more references to literature (Revision 1 PP17-18 L264-294). The redundancy pointed out is due to the structure where we focus one aspect of our observation at a time to which we give genealogical interpretation, but we believe that this structure keeps coherence and readability of each paragraph. To make the purpose of each subsection clearer, we added three subheadings under “Distinct patterns of genetic variation at low-recombining regions: Genealogical interpretations” in Revision 1: “Genealogical noise, genealogical bias, and mutational noise” (revision 1 P17 L266), “Species-wide low-recombining regions” (revision 1 P18 L295), and “Population-specific low-recombining regions” (revision 1 P19 L336).