Dear Dr. Novembre,

Thank you for assigning three highly qualified reviewers to our manuscript (ID# GENETICS/2018/300875), “Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity”, and for providing a small extension (granted by Ruth Isaacson on August 10, 2018) to submit the revision to our manuscript. The reviewers were overall positive about our manuscript, but had several comments and suggestions that would improve our work. We have carefully considered each of the reviewers’ suggestions, and made every effort to address each of them below, incorporating the majority into our revised manuscript.

Most prominently, we have augmented our discussion of the strengths and limitations of our methods at various points in the manuscript. We examine the performance of each method under models of population admixture and substructure, finding that we are robust to these effects in all but the most extreme circumstances. Furthermore, we explore a separate ABC classification strategy from our existing one in which we assign the most probable number of sweeping haplotypes to each selection candidate, and from this obtain inferences that support and complement previous results. Moreover, we have attempted in figures and in the main text to make our arguments more accessible to read and understand. Finally, to reduce the length of our manuscript, we cut our treatment on recent balancing selection, overhauled our admixture experiments, and attempted to further consolidate paragraphs throughout. However, because our extensive revisions included a number of new experiments, we were unable to make the manuscript substantially smaller than our initial submission.

In the following document, reviewer comments are indicated in gray font, while our responses to each comment follow in indented larger black font. We have marked all changes in our manuscript in red for ease of review, and also assigned line numbers for convenience.

We hope you agree that we have addressed all reviewer concerns, and appreciate you considering our work for publication in *GENETICS*.

Sincerely,

Alexandre M. Harris, Nandita R. Garud, and Michael DeGiorgio

Editor’s comments

Three experts in the field have reviewed your manuscript, and I have read it as well, and I am pleased to tell you that it is potentially suitable for publication in GENETICS. However, the reviewers have comments and concerns that need to be addressed in a revised manuscript. You can read their reviews at the end of this email.

If you submit a revised manuscript please include a response to each of the reviewers' comments. It is most important that you address the following in a revised manuscript: 1) The second and third reviewer both had concerns regarding the impact of cryptic population structure and whether the resulting departures from Hardy Weinberg proportions would impact the method. In addition, a related issue is whether admixture can create false-positive soft-sweep signatures. Doing a few illustrative simulations to address whether these concerns are substantiated and providing cautionary text in the discussion in these regards should be sufficient. 2) The third reviewer also recommends removing redundancy in the manuscript, which we agree would be helpful.

I look forward to receiving a revised manuscript. I expect it could be submitted within 60 days, but please let me know if you think you will need more time to complete the revision.

> We thank the editor for handling and reading our manuscript, and considering it for publication in *GENETICS*. We appreciate the editor’s and the reviewers’ interest in our work, and have undertaken numerous updates to the content of our manuscript, following the helpful suggestions of our three reviewers. Due to the addition of a number of new analyses, we were unable to substantially shorten the main text, though we did remove our balancing selection results to make space for our new population structure results, overhauled the admixture experiments, and condensed sentences and streamlined text when possible. We have also included more supplementary figures and tables to support our updated claims.

Reviewer #1 comments

In a recent paper, Garud et al. (2015) developed a series of simple and intuitive statistics (denoted H12 and H2/H1) to test for and classify hard and soft selective sweeps using phased DNA sequence data. The current manuscript extends Garud's approach by presenting and evaluating a new set of statistics (G12, G123 and G2/G1) that can be calculated from unphased data. This is an important extension given the ubiquity of unphased data (particularly when working with non-model systems). The manuscript is very well written and was a pleasure to read. Simulations are used to explore the efficacy of the method across a reasonable range of conditions and the method is also successfully applied to human genetic data. Key limitations of the method are mostly addressed. In total, I think this will be a useful tool and one that can readily be applied to many data sets. Nonetheless, I have a few (mostly minor) suggestions for revising the paper.

> We thank the reviewer for their thorough evaluation of our manuscript, and their positive remarks about our work. We have implemented the reviewer's suggestions in our revised manuscript.

1. Hard vs. soft sweeps. In general, I found the presentation of hard vs. soft sweeps in the introduction very lucid. With that said (and as is commonly done in the literature), hard vs. soft sweeps were given as binary categories, and (importantly) this categorical thinking extends to the simulations/classification associated with the method (more on the latter below). In some ways it seems more useful to think of the number of haplotypes contributing to a sweep than to always dichotomize this as hard (k=1) and soft (k>1). For example, a hard sweep (k=1) and a soft sweep with k = 2 (or k=3) are in some ways more similar than either are to a soft sweep with k = 20 (or something like that). I don't want to make too big of a deal about this and I realize the way things are described is consistent with current usage and emphasis in the literature, but it might be nice to have a sentence pointing out that this is really a continuum.

> This is an interesting point, and we agree that it makes sense to consider sweeps as a continuum of softness (*i.e.*, number of distinct sweeping haplotypes) rather than as a binary characterization of hard or soft. We have updated some phrasings throughout to acknowledge this point (for example, lines 9-10 of page 3). Additionally, we carried out the experiment proposed in comment 3 below, and generated important and helpful insights from our analysis of the posterior distribution of the number of distinct sweeping haplotypes *k* in the context of an updated ABC approach. We believe these experiments also highlight the concept of sweeps as a continuum. (Page 3, lines 9-10; page 12, lines 16-31; pages 13-15; page 19, lines 4-7; pages 19-21; page 26, lines 29-32; page 30, lines 11-17; page 32, lines 1-3; Figures S8-S10)

2. Simulating soft sweeps. Simulations for soft sweeps are described at the top of page 24. This is mostly clear, but there is one point that needs clarification, and that could be an issue. The manuscript states that selected mutations are introduced in k = 2,...,32 haplotypes (out of N = 1000). I can interpret this two ways. Considering as an example k = 2, this could mean that 2 of the 1000 gene copies in the

population would get the mutation or that two distinct haplotypes will be drawn from them and that all copies of those haplotypes would have the mutation. These are of course the same thing if every gene copy (sequence) in the population is unique, but otherwise they are not. The former would simulate a sudden pulse of new mutations (the recurrent mutation version of soft sweeps, but oddly with all of the mutations happening at once), whereas the latter is closer to what one might expect with actual standing genetic variation. Please clarify and justify the way soft sweeps were simulated.

> We understand that our original phrasing was somewhat ambiguous, and we have updated it in the results and methods. Our protocol involved drawing *k* haplotypes without replacement at the time of selection, and adding the selected mutation *de novo* to these haplotypes. Thus, scenarios in which *k* scaled haplotypes carry the selected mutation have a scaled frequency *k*/*N* and correspond to scenarios in which the the frequency of the selected mutation is at 20*k*/*N* in unscaled simulations (scaling factor λ =20). We performed soft sweep simulations in this manner to keep the simulation protocol as similar as possible across experiments with different values of *k* (thus, only one variable changes between them). We have made edits throughout the manuscript indicating that we are drawing different, but not necessarily distinct haplotypes, keeping the protocol otherwise identical to that for hard sweeps. (Page 8, line 24; page 16, line 29; page 27, line 7; Figure S4)

3. The ABC approach. The justification for k = 3 or k = 5 for the ABC inference was not clear to me. I think it would generally be better to instead place a prior on k and use ABC to then estimate a posterior for k (instead of using Bayes factors). From this on could calculate the posterior probability of, e.g., k = 1 (a hard sweep) or k > 1 (a soft sweep), but also simply obtain credible intervals (of various sorts) on k. Then, even if you couldn't confidently infer k = 1 (for example) you might still be able to conclude that k was likely 1 or 2 (which is itself informative). This goes back in some ways to my earlier statement about the distinction between hard and soft sweeps, but even if one ignores that, the current choice of k would need to be justified.

> Our ABC approach in the original version of our manuscript broadly followed the approach of Garud *et al*. (2015), in that our aim was to demonstrate the ability of our approach to distinguish between a model of a hard sweep and models of soft sweeps. To this end, we presented the (G12,G2/G1), (G123,G2/G1), (H12,H2/H1), and (H123,H2/H1) values for comparisons between hard sweeps and two soft sweep models (*k*=3 and *k*=5), illustrating that we are able to resolve these. We chose *k*=3 and *k*=5 because our experiments indicated that the expected homozygosity methods could easily detect sweeps within this range of softness (Figures 4 and S3). We also decided to use these results in particular because we found that each scenario yields a distinct and informative signal profile of (H12,H2/H1) and (G123,G2/G1) values (Figures 5 and S7). Thus, our results show that our approaches can distinguish between hard and soft sweeps over much of the values we tested. Additionally, changing the comparison of hard sweeps from *k*=3 to *k*=5 changed the coloration of plots in Figures 5 and S7 in the manner we expected. Under *k*=5, the occupancy of soft-assigned

values was reduced and shifted them toward more intermediate H12, H123, G12, and G123 and larger H2/H1 and G2/G1 relative to *k*=3, and we believe that demonstrating this progression is informative.

We thank the reviewer for mentioning this interesting experiment about inferring the posterior distribution of *k*. We carried out the reviewer's suggested experiment and inferred the posterior distribution of *k*, where the prior on *k* was uniform with $k \in \{1, 2, \ldots, 16\}$ (Figures S8 and S10; Tables S3-S14). Our simulated results here are concordant with those for Bayes factors (BFs), showing that soft sweeps (inferred $k > 2$) occupy smaller values of H12, G12, H123, and G123 with larger values of H2/H1 and G2/G1 (Figure S8). Accordingly, probability density functions of all replicates grouped by inferred *k* (Figure S9) indicate a shift in the values of H12 and G123 from larger to smaller with increasing *k* and smaller to larger H2/H1 and G2/G1.

We found that most probable estimates of *k* within the top 40 candidates of all populations ranged from 1 to 5, and strong concordance existed between most probable *k* and our previous BF estimates. The reviewer's suggested experiment was quite helpful in providing an inference of *k* that matched observed patterns in the data, and allowed us to reduce the ambiguity associated with classifying sweeps as hard or soft. Accordingly, we updated the manuscript to include our inference of the most probable *k* for each top candidate of each population, including the credible interval. However, the credible intervals for *k* were wide, and so BFs were still useful in providing support for our classification of many top candidates.

4. Other minor comments:

- It might be nice to discuss the relative merits of attempting to phase data statistically and using the H statistics, versus simply using the G statistics proposed here.

In the first paragraph of the *Discussion*, we now briefly explore the merits of applying the MLG-based methods rather than attempting to phase haplotypes in situations where doing so may be prohibitive. We note that "Because phasing may be difficult or impossible given the resources available to a study system, while also not being error-free [Browning and Browning, 2011, O'Connell *et al*., 2014, Laver *et al*., 2016, Castel *et al*., 2016, Zhang *et al*., 2017], the importance of our MLG-based approach is apparent." and that results for haplotype and MLG data were congruent. We hope that we have provided sufficient justification to reinforce our claims. (Page 15, lines 30-32; page 16, lines 1-2)

- Uncertainty in genotypes is another common issue with DNA sequence data, particularly when dealing with non-model systems. This could be accounted for in the proposed method by generating estimates of the G statistics over the posterior distribution of genotype estimates. Maybe this is worth mentioning or trying; I think it would be of interest to many folks, particularly given how easy it would be.

> We agree with the reviewer that this is an important idea, and one that many researchers are likely to consider in their work. Though we have not carried out experiments accounting for genotype uncertainty, we have devoted consideration to this topic in our updated *Discussion* that we believe sufficiently highlights this point. (Page 18, lines 12-17)

- Panels C and D for figures 3 and 4 appear to have four colors/distinct lines. But if these are for the same simulations as panels A and B, shouldn't these have one line per frequency cutoff (Fig. 3; 10 lines) or k (Fig. 4, 5 lines)?

> We have made it more explicit that we only show spatial signal curves for the parameters under which we have the most power. Our intention is to emphasize the genomic signature that should be expected for sweeps that our approach can properly detect. (Figures 3, 4, S2, and S3)

Reviewer #3 comments

Harris and colleagues present a simple but potentially very useful extension of the H12 statistic. Although haplotype based statistics, such as H12 are interesting because has been shown to be able to distinguish between classic hard and soft sweeps, their application has so far been limited to a handful of human and Drosophila datasets, organisms in which fully phased, large re-sequence datasets exist. Harris et al develop a straightforward set of statistics (G12, G123) which is analogous to H12, but instead of being based on the expected heterozygosity of the most frequent classes of haplotypes, considers frequencies of diploid multilocus genotypes (MLG). The authors show using simulation that, perhaps surprisingly, G12 and G123 have very similar properties (power, ability to distinguish between and robustness) to H12. They also compare the relative performance of H12 and G12 statistics on human data from the 1000 genomes project and again show, convincingly I think, that although not requiring phase information, G12 gives very similar with a. I found this manuscript interesting and well written and believe that it would make a solid contribution to Genetics. I have a few suggestions/questions, which I hope are helpful in improving the MS:

We are grateful to the reviewer for their appreciation of our work and insightful comments. We hope that our revisions will satisfy the reviewer's concerns.

1) p.5 Harris at al's new statistic is based on the combined frequency of the three most common diploid multilocus genotypes (q1+q2+q3) and this is motivated as a direct analog to the H12 statistic which is based on the expected homozygosity of the two most common haplotypes (p1+p2). Harris et al note that "for a situation in which haplotypes X and Y are both at high frequency, diploid individuals of type XX YY and XY will exist at high frequency". However, as I understand from Figure 1, q1, q2 and q3 are solely defined via their relative frequency and are not in any way constrained to be compatible with the presence of the two underlying haplotypes. If so, it would be good to make this clearer somehow.

In our application of G12 and G123, we do not presently enforce a compatibility constraint between the number of high-frequency haplotypes and the number of high-frequency multilocus genotypes at the site of an outlying signature of elevated expected homozygosity. Our assumption is that under random mating, the presence of high-frequency haplotypes implies the presence of high-frequency multilocus genotypes. However, we find both in simulated (Figure S4, *k* = 2 example) and empirical results (Tables S3-S14, note *SYT1* and *RGS18* in YRI on page 20, lines 14-18) that the presence of a high-frequency haplotype alongside an intermediate-frequency haplotype (two elevated-frequency haplotypes) can yield two high-frequency multilocus genotypes rather than three. Thus, a constraint on the presence of a certain number of sweeping haplotypes may result in the omission of some candidates. We have now explicitly mentioned our lack of constraint in the manuscript (Page 5, lines 9-11; page 18, lines 17-22).

My questions are:

i) To what extent would it be possible and/or helpful to build in such a constraint?

We appreciate the reviewer’s interest in this point, but we have not found it necessary to constrain the outlying sites we pick up with G12 and G123 to fit a haplotype pattern. Our approach assumes that the presence of multilocus genotypes at high frequency implies a sweep, and we indeed find high concordance between candidate lists generated from haplotypes and multilocus genotypes, as well as in their Bayes factors and inferred most probable *k* (Tables S3-S14). In addition, because soft sweeps may have diverse effects on the number of observed high-frequency multilocus genotypes, we may not always see the same number of high-frequency multilocus genotypes for a given number of sweeping haplotypes. For example, as shown in our new Figure S4, a scenario of *k*=2 sweeping haplotypes may not produce three high-frequency MLGs, but instead may produce only two.

ii) Given two common haplotypes at roughly equal frequencies p1 and p2, we would expect the most common MLG to be at frequency q1=2*p1*p2 and runs of homozygosity at frequency q2=p1^2 and q3=p2^2. How often do we observe such runs of homozygosity in the human data in soft sweep regions?

Though we have not examined each of our top soft sweep candidates for concordance with the random mating assumption, those that we have examined suggest that candidates with high-frequency haplotypes yield the expected high-frequency multilocus genotypes (at frequencies comprising any of $2p_1p_2$, p_1^2, and p_2^2). Additionally, we emphasize that our outlying candidates list for haplotype and multilocus genotype data are highly congruent across each of the four human populations we examined, suggesting the soundness of our approach and its assumptions (Tables S3-S14). Furthermore, we performed the experiment with runs of homozygosity that the reviewer suggested in comment 3 (see below), and found that prevalence of intermediate length runs of homozygosity were positively correlated with G123 and negatively correlated with $\log_{10}$(Bayes factor). These runs of homozygosity experiments are summarized a new Table S2, and discussed on page 15, lines 5-20.

iii) It would be useful to make the connection with runs of homozygosity (and test based on those) more explicit here.

We agree with the reviewer that highlighting the connection between runs of homozygosity and selective sweeps would be useful. Previous research has indicated that short to intermediate runs of homozygosity (ROH) spanning tens to hundreds of kilobases are characteristic of recent sweeps (Pemberton *et al*. 2012, Blant *et al*., 2017). To this end, we have intersected our top candidates lists with the inferred coordinates of short to intermediate ROH from Blant *et al*. (2017). We found a significant positive correlation between G123 of our top candidates and class 4 (intermediate length) ROH, but no particular correlation of classes 2 and 3 (short lengths) among top sweep candidates. In addition, we found a significant negative correlation between $\log_{10}$(Bayes factor) and class 4 ROH because softer sweeps are expected not to produce ROH of this length (Table S2). These new results are discussed on page 15, lines 5-20. We note, however, that caution is warranted when interpreting patterns of class 4 ROH because these also emerge due to consanguineous mating and therefore not exclusively due to selective sweeps.

2) As Harris at al argue in the discussion, the new G stats are potentially most useful for exploring selection in organisms for which obtaining fully phased individual genomes (e.g. from parent offspring trios) is not possible. However, resequence datasets for those organisms are typically considerably smaller than the sample size of n=100 Harris at al consider in all their simulation based power tests. It would therefore be extremely useful to know how much power the new G stats have for smaller sample sizes, say 50 and 20 individuals which, although still large, are more typical for non-human studies.

We agree with the reviewer that demonstrating the application of G12 and G123 to smaller sample sizes is important. We have included results demonstrating the power of all methods for simulated sample sizes of n=25 diploids and show that we maintain comparable power to what we achieved for the larger simulated samples of n=100 diploids when detecting sweeps (Figure S19). We emphasize that the expected homozygosity methods still perform well for samples of at least 25 diploids, but that power is likely to decrease for smaller sample sizes. That is, enough diversity must be captured in the sample, and increasing the sample size increases the captured diversity (Pennings and Hermisson, 2006a [Soft Sweeps II]). Additionally, our approaches require substantially larger sample sizes to properly infer the number of sweeping haplotypes. This is because distinguishing between two classes of sweeps requires the detection of a more subtle signal than simply distinguishing selection from neutrality, necessitating more sequenced individuals. (Page 17, lines 20-27)

3) The set-up of the comparison of G and H stats on the Human data is clever: Harris et al compare the G and H on exactly the same phased data and reconstitute diploid MLGs by combining the two haplotyped of each individual. It would be very interesting to know how much more similar G123 trajectories and outlier sets based on them are to their H12 analogs when diploid MLG are constructed at random. This would allow to quantify the extent to which any differences between G and H stats are due to departures from HWE.

This is an interesting experiment. We performed this experiment and believe that it lends support to our approach, emphasizing that it may be unnecessary to make further assumptions in the application of G12 and G123 to empirical data. The lists of outlying candidates when multilocus genotypes were randomly constructed greatly resembled the lists emerging from non-random pairings, but with the ranks of candidates, as well as Bayes factors and most probable k changing slightly, especially toward the bottom of the lists (*i.e.*, least significant candidates). Additionally, the outlying candidates found with non-random MLGs substantially overlapped the candidates found with haplotypes. (Tables S3-S14)

4) A potential confounding scenario not considered here is population structure. Structure causes deviations from HWE and depending on the sampling scheme may lead to two haplotypes at intermediate and so generate false positives. It would be very useful to include some simulations at least for a simple structured population to show how sensitive the new G stats are to structure when sampling is clustered (e.g. two sampled demes in a symmetric island model with a large number of unsampled demes).

We agree that population structure deserves greater attention, and accordingly performed the requested experiments demonstrating the application of the expected homozygosity statistics to a symmetric island migration model consisting of six demes, of which two are sampled (Figure S21A). We found that population structure, with low between-deme migration rates, yielded at most only somewhat inflated values of H12 and G123 in the absence of a sweep, not yielding values comparable to a recent strong hard sweep except for the smallest rates of inter-deme migration. (Figure S22)

- p. 5 bottom "simulations following human parameters", reword; I know what you mean but it sounds odd.

We have changed instances of this phrasing throughout the manuscript ("parameters compatible with our recent understanding of human demographic history," "population-genetic parameters inferred for human data," etc.) and believe that the relevant sentences now flow more naturally. (Page 2, lines 17-18; page 5, line 29; page 22, lines 19-20; page 25, line 15)

- p. 13 "... while NNT was significant for G12 only" How can this be given that you've shown that G stats have slightly lower power than the H equivalents in the simulation study. Is this because G stats are sensitive to departures from HWE?

There was a minor error in our initial assessment of candidate significance. This error has now been fixed, and all relevant results have been updated. *NNT* is now no longer significant for analysis using any method, though it is still a top 10 sweep candidate in YRI. Our results have also remained mostly identical to previously, with the order of candidates shifting slightly. Thus, we do not believe that our assumption of random mating results in any of the minor discrepancies between top 40 candidate lists. (Tables S3-S14; Figures S11-S18)

- throughout the paper: "method power", and "method sensitivity" is grammatically odd: either drop "method" or use articles.

We have changed all instances of this phrasing throughout the manuscript.

Reviewer #4 comments

# Summary

In a very well-written manuscript that seems technically flawless, Harris et al. present an approach for detecting and classifying selective sweeps from unphased multilocus genotypes (MLGs). The method extends the haplotype-homozygosity based approach of Garud et al. (2015) multilocus genotype (MLG) identity. Although there are now several methods for detecting selective sweeps based on linkage disequilibrium or extended haplotype homozygosity, all of them except for the one by Garud et al. (2015; NR Garud being a coauthor of the present manuscript, too) lack the ability to distinguish between hard and soft sweeps. Garud et al. (2015) have shown that the ratio of the haplotype homozygosity for all but the most frequent haplotype (H2) to the overall haplotype homozygosity (H1) has power to delineate soft and hard selective sweeps. This discrimination power depends on a statistic H12 (H123) that represents the haplotype homozygosity after the two (three) most frequent haplotypes have been pooled, and which has power to detect but not categorise sweeps.

The significance of the approach by Harris et al. is hat the MLG analoga to the H-statistics, i.e. G1, G2, G12(3), can be computed from non-phased data. This is a major asset given that phased genomic data are still scarcely available beyond a small set of model organisms, and so extends the applicability of the framework substantially. Yet, the conceptual innovation and originality of the manuscript appears to be within narrow bounds given the previous work by Garud et al. (2015). Harris et al. do make an important contribution though by thoroughly testing the robustness of their MLG homozygosity approach to confounding factors including population bottlenecks, population expansions, balancing selection, background selection, and the way missing data are deatl with. The authors perform forward simulations to measure the power to detect sweeps under these various scenarios, and to assess the ability to discriminate between soft and hard sweeps. Harris et al. illustrate their results in a visually very appealing way. The authors also apply their approach to phased haplotype data from four populations of the 1000 Genomes project. They corroborate previously identified sweeps, but also indentify novel ones. A pronounced difference between African (YRI) and non-African (CEU, GIH, CHB) populations in the relative abundance of identified hard vs. soft sweeps calls for an explanation and should be discussed to some larger extent.

Harris et al. find that the MLG identity statistics have generally high power to detect sweeps, and to classify them as soft or hard in substantial parts of the space of the MLG statistics. In particular, the methods is fairly robust against deviations from a constant-size demographic history and to background selection. Generally, detection power is highest for strong and recent sweeps, as expected. However, balancing selection can severely interfere with the detection power and lead to a high false-postitive rate. The authors also study the effect of admixture masking some of the sweep signal, and find that if single-pulse admixture proportions are sufficiently low (< 0.3), then their MLG statistics retain power to detect a hard sweep.

I think that the manuscript is of high interest and very nicely done. I did have a few concerns (see below) related to the robustness and biological scope, as well as to the assumption of Hardy-Weinberg equilibrium. Overall, I am undecided as to whether in its current state the manuscript qualifies for publication in GENETICS, since, from a technical point of view, it is to a large extent of an extending, rather than innovative, nature. I think the latter concern could be removed if some of the major concerns mentioned below (false positives under admixture, robustness to non-random mating, visualisation of simulated likelihood surfaces) were included, as these points would on their own add substantial biological insight and guide future applications of the method. That said, the manuscript is quite long already, and I suggest that the authors make another effort to shorten it. I feel there is quite some scope for condending the main text, especially by removing redundancy in the description of the methods or moving some of it to the supporting information.

We thank the reviewer for their thoroughness in evaluating our manuscript, and appreciate their praise of our work. We have implemented the suggested changes to the best of our ability in the context of our updated and new results.

# Major concerns

1. The principal reasoning underlying the new approach by Harris et al. is that the most frequent haplotypes yield the most frequent (homozygous) MLGs. However, this holds only under the assumption of random mating. In certain organisms, such as plants that (partially) self, or in populations with a cryptic population structure, mating is not random. Natural selection, too, may lead to a distortion of Hardy-Weinberg proportions at least temporarilly. Therefore, it would seem appropriate if the authors tested the robustnes of their method to deviations from the assumption of random mating (e.g. due to partial selfing or cryptic population structure).

We performed an experiment to determine the effect of population substructure on the value of the expected homozygosity statistics, simulating a symmetric island model with six demes, of which two are sampled (Figure S21A), and found that inflation of the expected homozygosity statistics can occur in situations where the migration rate between demes is very low. This inflation is not comparable to that generated under recent and strong sweeps on few haplotypes in unstructured populations. Thus, violations of the random mating assumption may reduce only our power to detect weaker, softer, and more ancient sweeps. (Figure S22)

2. While the authors study the effect of admixture masking an existing hard sweep, they leave unaddressed the question of whether and to what extent admixture alone can create false-positive (soft-)sweep signatures. I think this should be addressed with another set of simulations focussed on establishing the effect on detection and classification power as a function of admixture strength and time. See also major concern **5.** below.

We agree that this is an important consideration and we ran comprehensive, updated admixture simulations in which a pulse of admixture from an unsampled donor population occurred in the history of the sampled population before the time of sampling (Figure S21B). We found that the only scenario in which admixture potentially produces a spurious signal of a soft sweep (with elevated G123 or H12 and elevated G2/G1 or H2/H1) is if the donor population has a small effective size and admixes at a high proportion. Outside of this narrow condition, admixture alone does not create false signatures of soft sweeps. (Figures S23 and S24)

Where is it discussed ?

3. The authors do not report how the power of their approach to discriminate between hard and soft sweeps relates to the frequency with which test combinations of G12 (G123) and G2/G1 are encountered in their simulations, given their prior believes in model parameters. In other words, it is unclear how often biological systems occupy the summary-statistics space for which detecion and discrimination power are highest. Figures 5, 6, and S6 are very nice and provide much insight. It would be desirable, though, that the authors also showed the joint distribution of the simulated statistics, e.g. of the (G12, G2/G1) values in the case of Figure 5 A and B, separately for hard and soft sweeps. This would reveal how the ability to distinguish between the two types of sweeps relates to the ferquency at which the test-value combinations are to be expected. If Bayes factors are high or low only in situations that have a very low absolute frequency, then this would somewhat reduce the encouraging results shown in terms of method power and Bayes factors only.

We agree with the reviewer that the point of practicality is an important one, and we have strived to demonstrate it in our work. We believe Figure 6 appropriately illustrates that empirical data is likely to exist within the range of G123 and G2/G1 values for which resolution is greatest. This range corresponds to intermediate values of G123 and a wide range of small to large values of G2/G1. To address the reviewer's request, we have augmented Figures 5, 6, and S7 (formerly S6) with informative colored bars indicating the number of hard and soft sweep simulations falling within the observed range of values. Our updated results indicate that the majority of replicates for both hard and soft sweeps occur at smaller values of H12, H123, G12, or G123, with a somewhat more even distribution of H2/H1 or G2/G1. Thus, the computation of Bayes factors for our empirical top candidates relied on thousands of nearby observations to classify sweeps as hard or soft. (Figures 5, 6, and S7)

Moreover, we have now included a new set of experiments in which we compute the most probable number of sweeping haplotypes *k* for paired (H12, H2/H1), (H123, H2/H1), (G12, G2/G1), and (G123, G2/G1) values. These new experiments indicate the space of values for which certain numbers of sweeping haplotypes are most prevalent, and should provide some information about the joint distribution of paired values (e.g., H12 and H2/H1) for different sweep softness. (Figures S8-S10)

4. Related to the previous concern, the reader's intuition of what makes the MLG identity statistics powerful statistics could be increased if the authors visualised the simulated (i.e. empirical) probability mass function of the G statistics as a function of the parameters (as a proxy of the joint likeihood function). This would approximate the marginal distribution of each of the statistics (e.g. G123, or

G2/G1) as a joint function of the start time (t) and strength (s) of the sweep, and could be done for a small set of meaningful combinations of f and k.

Following the reviewer's helpful suggestion, we generated the probability density functions of H12, H2/H1, G123, and G2/G1 for simulated replicates of selective sweep scenarios across $k \in \{1, 2, \ldots, 16\}$. We indeed observe that across the spectrum of *k* values we examined, increasing the number of sweeping haplotypes results in a shift of H12 and G123 toward smaller values, and a shift of H2/H1 and G2/G1 values toward larger values, with either extreme showing a markedly different distribution. (Figure S9)

5. The applications to a subset of the 1000 Genomes data reveal the striking pattern that the YRI show a much larger relative proportion of soft vs hard sweeps, as compared to the CEU, GHI, and CHB populations. I was wondering if, besides the out-of-Africa bottneleck(s) and subsequent adaptation explaining a potential excess of recent hard sweeps in non-African populations, underappreciated admixture in YRI (e.g. Busby et al. 2016, eLife, doi: 10.7554/eLife.15266) could have created some apparent soft-seep signals, and so be part of the explanation of this striking difference. A detailed investigation of this question seems beyond the scope of the paper. However, I think that addressing major concern **2.** above will provide some insight, and the issue deserves some more discussion in the main text.

Upon further evaluation of our empirical results as indicated by the reviewer, we observe that the YRI population generally has more candidate soft sweeps than do the CHB and GIH populations in the list of top 40 candidates, though each of these populations yielded noticeably more candidate soft sweeps than did the CEU population. However, across all four human populations, there were more candidate hard sweeps than soft sweeps in each population. We judged this using our updated ABC approach, from which we assigned a most probable underlying value of *k* from the posterior distribution of five million replicates with *k* drawn uniformly at random from a prior distribution (see updates to *Materials and Methods*). Across the haplotype and MLG experiments, for YRI we found that between 35% and 47.5% of candidate sweeps were soft (depending on method and data type), whereas between 22.5% and 37.5% of candidate sweeps in GIH, and 22.5% to 30% of candidate sweeps in CHB were inferred to be soft. In contrast, only 10% to 17.5% of candidate sweeps in CEU were soft. (Tables S3-S14)

In conjunction with our admixture experiment in response to the reviewer's comment 2, we therefore do not believe that the admixture described in Busby *et al*. (2016) accounts for the proportion of soft sweeps observed as top candidates in YRI. We found that only admixture from a donor population whose effective size is an order of magnitude or so smaller than the target (recipient) population can spuriously raise both G123 (H12) and G2/G1 (H2/H1) values. Moreover, this would affect the whole genome, which is not what we observe (Figures S13 and S14). Instead, we believe it to be more likely that our results reflect the higher occupancy of large Bayes factors greater than one for paired (G123, G2/G1) values for YRI than for the other populations (Figure 6). That is, the large effective size of YRI makes it easier to detect the normally weaker signal of soft sweeps.

# Minor comments

Abbreviations used:
- C: comment
- Q: question
- S: suggested change [a -> b means replace a by b]
- R: requested change [a -> b means replace a by b]

## Abstract

- [Second-to-last sentence] S: "...based on human parameters.." -> "...based on parameters compatible with our recent understanding of human demography..."

We have changed various instances of this phrasing throughout the manuscript.

## Introduction

- [p. 3, l. 3] S: "...signatures, hard sweeps and soft sweeps,..." -> "...signatures, those of hard sweeps and soft sweeps, respectively,..." [I think the formulation depends on whether a sweep is considered a process or a pattern; I favour the former]

We did not implement this suggestion because the structure of the sentence in question has changed. (Page 3, lines 4-6)

- [p. 3, l. 21] S: "selected allele" -> "one of the selected alleles"

We retained our original phrasing because we are referring to the selected allele at a biallelic site, existing on multiple haplotypic backgrounds and reaching fixation. (Page 3, line 22)

- [p. 4, l. 1] R: Fix citation style for Chen et al. (2015)

We have corrected the citation for Chen *et al*. (2015). (Page 4, line 1)

- [p. 4, l. 7] S: "...computed as expected..." -> "...computed as the expected..."

We have incorporated this suggested change. (Page 4, line 6)

- [p. 4, l. 17] S: "...is expected haplotype..." -> "...is the expected haplotype..."

We have incorporated this suggested change. (Page 4, line 16)

- [p. 4, l. 25] S: "...MLGs are a single string representing a..." -> "MLGs are single strings representing a..."

We have incorporated this suggested change. (Page 4, line 25)

## Results

- [p. 8, l. 31] R: "...produces..." -> "...produce..."

We did not implement this suggestion because the structure of the sentence in question has changed. (Page 9, line 1)

- [p. 11, l. 23] C: Here and throughout the rest of the paper, I found the use of "parameter space" inappropriate, as the G and H statistica are not parameters, but summary statistics. I suggest using a different term.

We have ceased referring to “parameter space” and now use more appropriate terminology throughout. As examples, we now tend to use phrases like “(H12,H2/H1) and (G123,G2/G1) values” or “paired (G123,G2/G1) values”.

- [p. 12, l. 31] R: Fix citation style for Huber et al. (2016)

We have corrected the citation for Huber *et al*. (2016). (Page 13, line 19)

## Discussion

- [p. 14, l. 28] S: "...sweeps, resulting in a..." -> "...sweeps, which results in a..."

We have incorporated this suggested change. (Page 16, line 7)

- [p. 15, l. 1] S: "...further back in time..." -> "...if they started far enough back in time..."

We have incorporated this suggested change. (Page 16, line 11)

- [p. 15, l. 4] R: "...allele decreases." -> "...allele decreases due to recombination."

We have incorporated this requested change. (Page 16, line 15)

- [p. 15, l. 6-9] C: The flip side perhaps worth mentioning is that it is difficult to detect and classify weak positive selection.

We have acknowledged this point briefly in the *Discussion* where the reviewer has requested it. (Page 16, line 18)

- [p. 16, l. 26-30] S: Remind the reader of how the Bayes factor was designed, i.e. that it is P[soft]/P[hard], not the inverse.

We have incorporated this suggested change. (Page 18, line 32)

- [p. 19, l. 25-28] C: It seems very plausible that G12 and G123 (or H12 and H123) have weak power to distinguish an almost complete hard sweep from balancing selection if h approaches 1, as the latter is then close to directional selection. However, I was wondering if the authors could add G2/G1 into the mix here, and see if the combination of G12 and G2/G1 has more power to differentiate between a hard sweep and balancing selection than has G12 (G123) alone.

We agree that it is interesting to see if G2/G1 together with G123 could distinguish hard sweeps from recent heterozygote advantage for small dominance coefficients, similar to the Bayes factor analyses for distinguishing between soft and hard sweeps. However, we unfortunately found that G2/G1 did not play a large role here, and instead G123 was the major factor. That is, scenarios with large G123 were more often classified as hard sweep than recent balancing selection, and the extra G2/G1 dimension was unable to help separate the space for large G123 values. However, we have removed the extensive discussion of balancing selection from our revisions in an attempt to condense the manuscript, and because in many cases signatures of recent and strong heterozygote advantage will be indistinguishable and lead to the same outcome as a selective sweep.

- [p. 19, l. 32] S: "...reduction in nucleotide diversity,..." -> "...reduction in nucleotide diversity and a distortion of the site-frequency spectrum,..."

We have incorporated this suggested change. (Page 22, line 15)

- [p. 20, l. 12-14] C: Besides obscuring the true signal of a sweep, admixture (from a ghost population) may also create a false signal of a sweep, and so I think the authors should assess the false-positive rate under simulations of a neutral admixture scenario, with varying admixture times and proportions. See major concern **2.** above.

We agree that this is scenario is important to consider. We did not specifically assess the false positive rate under neutral admixture settings, but in Figures S23 and S24 we show that simultaneously elevated H12 and H2/H1 (Figure S23), as well as G123 and G2/G1 (Figure S24), can occur for small donor effective size and sufficiently large admixture proportion. However, this admixture scenario was the only setting for which we would observe a false sweep, which would also be consistent with a soft sweep.

- [p. 21, l. 25] S: Based on the results shown, it seems bold to claim that the method by Harris et al. provides a means of distinguishing sweeps from balancing selection without saying that this distinction is restricted to a set of conditions that is hard to verify in practice. It would seem more conservative to claim that the method can differentiate between selective sweeps and background selection.

> We agree that it was necessary to rephrase this assertion considering the wide range of scenarios under which recent sweeps and recent strong heterozygote advantage produce similar genomic signatures. Because we have removed our discussion of balancing selection to condense our manuscript, we have also removed this assertion.

- [p. 21, l. 26] S: "...differentiation lends itself well for use as a..." -> "...differentiation motivates the use of MLG identity statistics as a..."

> We have incorporated this suggested change. (Page 24, line 22)

## Materials and Methods

- [p. 22, l. 7] S: Add recombination as a process simulated by SLiM 2

> We have incorporated this suggested change. (Page 25, line 7)

- [p. 23, l. 11] S: "...heterozygote advantage selection..." -> "...heterozygote-advantage selection..."

> We did not implement this suggestion because we have omitted balancing selection from our updated manuscript to further condense it.

- [p. 23, l. 16] R: Insert a comma after both occurrences of "G2/G1" in this sentence.

> We did not implement this suggestion because the structure of the sentence in question has slightly changed. (Page 26, line 15)

- [p. 23, l. 25-26] R: Please specify the base of the logarithming prior scales.

> We have incorporated this requested change. (Page 26, lines 26-27)

- [p. 24, l. 3] C: The authors seem to use $N$ to denote both the haploid and diploid population size, depending on the context. As this is confusing, I suggest to change this.

> We agree that our notation needs to be rigorous and consistent. We have updated our notation to define the diploid population size as *N* and the haploid as 2*N* throughout.

- [p. 24, l. 25-28] S: "Balancing selection simulations..." -> "Simulations of balancing selection...". C: To be more precise, I suggest using "overdominant directional selection" to describe what the authors currently call "balancing selection". At least, this specification should be made once, if the authors want to stick to the term "balancing selection" in the remainder of the paper.

We have removed this from the manuscript because we have omitted balancing selection experiments to condense our manuscript.

- [p. 25, l. 1] S: "...our single background selection scenario..." -> "...our single background-selection scenario..."

We did not implement this suggestion because the structure of the sentence in question has changed. (Page 28, line 3)

- [p. 26, l. 10] S: "...spatial signature..." -> "...genomic signature..."

We have changed instances of “spatial signature” to “genomic signature” where we refer to empirical results or general application of the expected homozygosity methods (Page 5, line 17; page 15, line 17; page 24, line 20), and leave references to the “spatial signature” along simulated chromosomes as-is (Page 8, lines 1 and 21; page 9, line 18; Figures 3, 4, S2, and S3).

- [p. 26, l. 21-22] S: "...under a constant population size demographic history..." -> "...under the demographic scenario of a constant population size..."

We did not implement this suggestion because the structure of the sentence in question has changed. (Page 28, lines 6-7)

- [p. 28, l. 1-5] R: Please split this sentence, it seems too long. Q: How important is this additional filter for mappability and alignability?

We have updated the sentence to make it into two shorter sentences. The CRG100 filter for mappability and alignability is important because it allows us to remove analysis windows in which variant calls may be unreliable. Figures S11-S18 demonstrate that filtered windows generally cluster near the telomeres and centromeres of chromosomes, where repetitive sequences of low diversity may yield elevated values of H12, G12, H123, and G123 in the absence of a sweep. (Page 31, lines 9-12)

## References
- [p. 35] S: "GENETICS" -> "Genetics" in reference to Loh et al (2013)

We no longer cite this paper in our current manuscript.

- [p. 36] C: Volume (and issue) missing for reference to Muhlfeld et al. (2009)

We no longer cite this paper in our current manuscript.

## Figures
- [Figure 2] S: I suggest the authors add an illustration of the admixture model (this could also be a figure in the supporting information).

We have added a supplementary figure (Figure S21) illustrating the specifics of our admixture (Figure S21B) and symmetric island migration (Figure S21A) models.

## Supporting Material
- [Figure S4] R: Please remind the reader of the selection coefficient used for the results shown.

We have incorporated this requested change. (Figure S5)

- [Figures S7 and S8] C: For Chromosome 2, it is not clear if all the gene names refer to the same peak (around 136 Mb) unless one also considers Figure 7. Q: What is the difference between the black and grey lines? C: This is a detail, but it appears as if the visualisation linearly interpolates across gaps due to missing information, which does not seem desirable.

In the CEU population, the labeled genes on chromosome 2 are all found within the signal cluster at position approximately 137 Mb. We have updated all figure legends of this series (Figures S11-S18) to indicate that the gray lines represent the value of G12 or G123 from filtered windows following the application of the CRG100 filter. We have also removed instances of improper linear interpolation between true signal locations.

- [Figures S19 and S20]: R: Please make the legend labels more self-explanatory. In the caption, replace "...to detect sweeps." by "...to detect hard sweeps.".

We have substantially revised these figures, and the relevant phrasing has changed. (Figures S23 and S24)