Red: needs to be changed in text
Green has been changed

NB note 7th May - I added some responses, but still use “branch” rather than “edge” At some point, one person needs to go through making usage consistent.

Editor Comments to Author:

Three reviewers have provided thorough assessments on this opinion manuscript. All authors are in agreement regarding the value of this contribution. I concur with this assessment. Additionally, as R1 highlights, the manuscript is timely given the increased reliance on long read data. However, all three reviewers also bring up important points that should be addressed before submitting a revision. While I agree with all points of criticism, I will highlight a few of them here. First, more background is needed on theoretical population genetic ideas that are relevant in this context (R2). In providing this information, the authors should aim to improve the readability of their paper and making it accessible to the diverse readership of Molecular Ecology. Also, the paper could reach a broader audience if a worked example would be provided on how an empirical dataset would be used in this context (including currently available tools and limitations; see comments from R1, R2, and R3). For example, how exactly might this definition of haplotype blocks benefit studies of selective sweeps? More details on the simulations used here are needed (R1), as are clarifications of the figures (R2). Lastly, the manuscript would benefit from increased discussion on the tradeoffs between the feasibility of analyses for realistic datasets, and how informative and analytically tractable results are likely to be (R3). I congratulate the authors for this contribution, and look forward to reading a revised version.

Commented [1]: elaborate in the practical section

Commented [2]: Incorporate Arka's analysis

<we can write a response to this comment at the very end>

<if we change branches to edges we might write a common note here>

Reviewer Comments to Author:

Reviewer: 1

Comments to the Author
Shipilina, Stankowski, Paul, Chan, and Barton make the case for using haplotype blocks, rather than SNPs, as the core focus of locus-specific population genomic studies. They argue that technological advances that enable experimental phasing, together with computational tools for ancestral recombination graph (ARG) inference, should allow studies to track haplotype blocks rather than individual mutations, and that this change in focus could help address shortcomings of current methods for inferring selection, population structure, and gene flow. The authors then propose a definition for haplotype blocks based on ARG topology and explore properties of haplotype blocks, thus defined, using simulated data under neutral evolution as well as a selective sweep.

This manuscript arrives at a time when big changes in population genomic inference are likely, due to advances in both data collection (i.e. long read sequencing) and data analysis (i.e. scalable ARG or tree sequence inference). Furthermore, the suggestion to move away from SNPs as the core focus of studies is a good one (the authors could cite the Ebert 2021 paper in Science that found a large indel in an intron of LCT that may be the causative mutation behind lactase persistence, rather than the previously reported causal SNP). This paper also provides some useful figures for understanding and visualizing properties of the ARG, which will assist many readers new to ARG literature. The 3D visualizations of haplotype blocks together with the branches they arose on are particularly helpful. I have several broad concerns about the manuscript as written, however.

The core problem this paper seeks to solve – confusing definitions of haplotype blocks – seems incompletely addressed, as the proposed definition depends on a choice of branch in the ARG. This allows for there to be many different overlapping haplotype blocks, some present in the same genomes, at any given locus. Without an inferred ARG, any given branch could still manifest as a set of SNPs in perfect LD, making it unclear why the ARG is necessary. The authors seem to suggest that one benefit knowledge of the ARG provides over SNP data alone is the ability to include regions of haplotype blocks that are not tagged by SNPs. Without informative SNPs, though, no ARG inference method could be relied on to get the history of these regions correct in the first place (this is alluded to on lines 224-227). And given the authors' concerns about potential biases in existing scalable ARG inference strategies (line 391-392, 837-840), it is unclear how the authors believe they should be identified. As it stands, choosing a branch in an inferred ARG does not seem much less arbitrary than choosing a set of linked SNPs that tags that branch (similar to, for example, S* from Plagnol and Wall 2006) – unless the authors, for example, mean to highlight that the ARG also tells us how far along the chromosome that haplotype persists. In addition to proposing this new definition for haplotype blocks, the manuscript should clearly state why and how ARG-based haplotype identification can benefit specific types of analyses.

Commented [3]: Our argument here could be that even though branch choice is random, by looking at different blocks we have some more informed decisions.

Commented [4]: I look at it differently: ideally, we should look at the whole set of branches, not choose one of them in an arbitrary way. It may be that when we look at that set, some stand out as being informative, but we don;t assume that in the paper.

Our paper is primarily about showing how “haplotype blocks” arise from the structure of the ARG, and showing how they can be clearly defined; we do not undertake the much larger task of developing practical inference methods from actual data. We focus on the ideal case where the ARG is known, and suggest ways to describe and visualise it. It may be that in practice, we can use SNP that are shared by a particular set of sampled genomes as surrogates for a branch in the ARG - but finding how accurate that is, compared with using an inferred ARG, is beyond our scope.

We do have in mind that the extent of a branch along the genome may gives important information, as the reviewer suggests. Again, finding how much extra information this gives, beyond the genealogy at the sweep (for example) is an open question.

Choosing a particular branch would indeed be arbitrary - rather, inferences should ideally be made from the full set of branches (that is, from the full ARG). In practice, we may consider the most substantial branches, tagged by sufficient SNP; how many are informative will depend on what question is asked.

The authors discuss selective sweeps – a topic of interest to many readers – but do not clearly explain how their definition of haplotype blocks will benefit future studies. Selective sweeps are often population-specific, but the example given shows an allele that swept to fixation in all lineages. In Figure 3D, the TMRCA (of all lineages?) is shown to be lower near the selected mutation, but in the case of population-specific selection, the TMRCA of all lineages is less likely to be a useful metric (the Rasmussen 2014 paper defined the “relative TMRCA halflife” as a way to identify regions where

Commented [5]: NB: Actually, examining even the simplest case of a hard sweep through a single population suggests that it will be hard to exploit additional information from haplotypes. Haplotypes may be useful in distinguishing amongst more complex models, including incomplete sweeps. However, I am sceptical that the solution is the invention of “better” statistics: rather, one needs a careful examination of each case.

Commented [6]: We chose the very simplest case of a hard sweep that fixes within a single population. There are of course many other scenarios - but the purpose of our paper is not to propose new methods for (say) detecting selective sweeps, but to clarify the more fundamental question of how we should define "haplotype blocks", and how we can use them to help visualise the ancestral structure.

some but not all clades might have an abnormally recent TMRCA). In the example given, the authors highlight ARG branches with large numbers of descendants as indicative of the simulated sweep, but they do not elaborate. A new, ARG-based selection statistic would be of use to many and could be explored further here (to complement, for example, those from Speidel 2019 and Stern 2019). As it stands, it is not clear exactly what the authors are proposing for the detection or study of selective sweeps, and the sentence on line 322-323 ends this discussion without making a clear point. Moving in the opposite direction – seeking to identify the selected site using the ARG topology – could be a useful exercise and could illustrate the power of using ARG-defined haplotype blocks in selection scans.

Commented [7]: We can discuss this

We again emphasise that we aim to find ways to better define and describe the ARG, rather than develop new statistical methods for specific estimation problems. We provide one illustration of the very simplest case, of a complete sweep, from one copy to fixation, in a single population. There are very many other scenarios, and we cannot investigate their consequences for the structure of the ARG here. Rather, we point out that on the one hand, the ARG contains much more information than is currently used, but on the other, there is considerable variability amongst different realisations, even under the simplest model.

The simulations are a large part of this manuscript but are not described in much detail in the text, aside from mentioning the infinite sites assumption and some of the parameters chosen. The supplementary files are very detailed and contain some beautiful visualizations, but a separate writeup (or reference to another publication in which the simulator is described) would be helpful. Readers might be interested, for example, how the simulator here compares to others with which they might be more familiar.

We added an introduction to the supplement which explains the simulation in detail, and this is summarised in the main text. The simulation itself is quite simple, using 12 functions, each with a few lines of code, to simulate ancestry conditional on the genotype at a selected locus. The bulk of the code defines tools that help describe the ancestry. The aim is to allow a detailed exploration of the ARG, which is rather different from other coalescent simulations such as msPrime.

Commented [8]: Nick

Commented [9]: I updated the SI, and added at the start a detailed explanation of the simulation algorithm. The first paragraph of that could go into the main text as a compact summary.

The authors mention that Relate and tsinfer, two recently-described ARG inference methods that scale to large data sets, both use the Li and Stephens haplotype-copying model and express concern that this could lead to bias (line 391-392, 837-840). This potential bias would arise from the existence of disjunct haplotype blocks, or specific haplotype blocks that exist in multiple, non-contiguous places in the genome, because the Li and Stephens HMM only considers the source/ancestral haplotype of the immediately preceding site when modeling a sampled haplotype. It seems to me, however, that the HMM could easily transition back to the same source haplotype again, if SNPs are encountered further down the chromosome in the sampled haplotype that again resemble that source haplotype closely enough. In other words, the model should work well enough if there are enough informative SNPs. If there are not enough observed SNPs, it is also hard to imagine any ARG inference strategy that could accurately assign ancestry in a given region. Do the authors have a suggestion on how to overcome this potential bias?

Commented [10]: Added

Commented [11]: There was another method recently published that I sent around. We should probably also mention it

Commented [12]: Discuss in the practical part

We have elaborated on the potential limitations of the Li and Stephens HMM, pointing out situations where the algorithm may be able to accurately detect transitions back to previous states. We have also pointed out that future work is needed to understand how problematic this would be for accurate inference of the ARG.

This bias could also be explored using the simulated data sets – since disjunct haplotype blocks have already been identified, one could infer ARGs using tsinfer and Relate on the simulated data and investigate how often disjunct haplotype blocks are successfully recovered. The authors could also explore the effects of the mutation/recombination rate ratio on the ability to identify such haplotype blocks. This could have implications for other studies – for example, there must be a tradeoff between the wide availability of reference data but low heterozygosity in humans, versus the scarcity of reference data but higher heterozygosity in some other species.

Commented [13]: This seems like a bit too much for me

Commented [14]: NB: This would be worth commenting on, but not a full-scale investigation

Commented [15]: yeas, highlighting this as something that we need to leave more about

Commented [16]: Worth adding, discussing?

Commented [17]: Yes, it's a simple & important point.

We have pointed out that more work is needed to evaluate and understand the performance of different methods for inferring the ARG under a variety of circumstances. However, this will take a substantial effort, well beyond the scope of our paper. (Brandt et al. (Genetics, 2022) consider the performance of these methods in recovering pairwise coalescence times, which is the simplest feature one can examine).

Finally, the conclusion (on lines 424-426) introduces the concept of graph genomes and alludes to problems with reference-guided assembly. Although structural variants are clearly an important and overlooked class of mutation, it is not clear how or whether the more complex classes of variants that graph genomes can uncover can be encoded into ARGs. Some structural variants, for example, could change the physical distance between neighboring mutations in some but not other genomes, which could affect the probability of recombination in the region and thus impair an ARG inference method's ability to infer recombination events. Additionally, methods for assigning ages to mutations depend on assumptions about mutation rates, which are well studied for SNPs and less so for other classes of variants. I have a similar concern with the statement about recombination events providing an alternative 'clock' to mutations (line 62-63), as it is not clear how or whether these two different pieces of information can be (or have been) reconciled to produce a single time estimate. More discussion of graph genomes and the complex variants they can describe would be welcome here, if the topic is raised.

Commented [18]: NB: I think that and ARG can still be defined - the recombination process simply becomes context-dependent.

Commented [19]: This is again a distraction. The ARG describes the ancestry; any kind of variant can be superimposed on it. Structural variants may be harder to use in inferring the ARG, and they may alter rates of recombination, but that does not affect the principles behind description of a population by the ARG.

There are two issues here. First, the ARG describes the ancestry of a set of genomes, which is defined independent of allelic state: all that is required is homology. Structural variants could change recombination rates, but the same is true of SNPs; this might raise practical difficulties, but doesn;t affect the fundamental principles. Second, mutation and recombination both act as "clocks" and both are used in existing methods, though often implicitly (e.g. in PSMC). Use of recombination as a 'clock' is perhaps most obvious for sharing of large blocks of genome, which indicates recent ancestry, at a time the inverse of the map length (e.g. Ralph & Coop, 2013).

Line-specific comments:

Line 50-56: Several problems with selection scans and GWAS are identified, but it is not made clear (i.e. in the following paragraph) how the new definition of haplotype blocks proposed here would address these problems. For example, it seems like a focus on haplotype blocks rather than SNPs could broaden the list of candidate causal mutations for a selective sweep. But selecting a variant from this list would likely require some kind of functional data.

Commented [20]: not really sure what to say here...

Commented [21]: Yes, and this is a key point: SNP are an imperfect reflection of the underlying haplotype structure, making it clear that it may be impossible in principle to locate causal mutations without functional information or, ultimately, experimental manipulation.

Line 106: Panel (A) indicator is missing from the beginning of the caption

This has been added

Line 137: This is shown later in Fig B1B, but mentioning that coalescences join lineages, while recombination events split them going back in time, might be helpful to readers.

We have made this more explicit in the text. This also relates to reviewer 2s request for clearer description and more background on population genetic processes.

Line 189-190: “The genome is divided into 34 non-recombining intervals…” – it is unclear to me whether this was a decision that went into the simulation, or a result observed after the simulation was run. More information about the simulations would help clarify this.

In addition to elaborating on the details of the simulation, we have expanded on this description to highlight that there are 34 regions that are defined by recombination events

Line 211: “(C & D) Examples of regions of blocks that, by chance, are revealed by mutations…” should be “are not revealed by mutations”

Done

Line 247-250: The concept of nested haplotype blocks is interesting. It also seems to exclude haplotype blocks defined by child branches – i.e. the block defined by branch i would exclude the block defined by branch ii. This was not immediately clear, though, and could be stated more explicitly.

Commented [22]: This is tricky - we should discuss. I mention the issue deep in the SI

Commented [23]: These kinds of issue become obvious as one works through an actual example....

We have discussed more clearly in relation to blocks rather than just branches. *Need to be more specific !*

Line 252: “a particular point on the map” uses different language than the rest of the sentence -- does this mean a particular site on the chromosome?

We have revised this to make our language use more consistent.

Line 254: “The rate of recombination is proportional to the branch length” – should this be changed to the “incidence of recombination,” since the underlying rate should depend on genomic span / local recombination rate, whereas branch length indicates the amount of time over which recombination events can occur?

Yes, this is correct–the length of time is proportional to the incidence or probability that recombination will occur. We have clarified this in the text.

Line 260-261: “In simulation, deep branches tend to be wider than expected…” – it is not clear exactly what the expectation was, or how much the simulated data differed from the expectation.

Commented [24]: Nick, I remember discussing this, but I can no longer remember why this was the case...

Commented [25]: I need to update the SI on this - it's tricky to explain...

Nick to address

Line 305-306: “9 of the most substantial branches are shown” – unclear here exactly what “substantial” means, and whether these are the only branches that passed the criteria that follow in parenthesis, or if these were chosen from that set. Could these branches also be highlighted somehow in panel B?

Commented [26]: Nick: caption for fig 4. how were significant branches identified here?

Commented [27]: It does say in the caption

We have clarified the details of how the branches were chosen: they are defined in the caption of Fig 4: “(These have more than 8 descendants, formed by coalescence more recently than the sweeping mutation, and have areas >0.5). “. We considered highlighting these specific branches in B but we think it makes more sense to highlight all of the branches that coalesce before the sweeping mutation rather than just specific branches. However, the coloured branches are now shown in the supplement.

Line 365: “hidden Markov model” appears twice

This has been fixed

Line 390-392: If haplotype blocks are poorly recovered by tsinfer and Relate, does ArgWeaver do a better job? Evidence of the proposed bias would be helpful, and ArgWeaver could be tested as a non-Li and Stephens-based alternative.

Commented [28]: arka, I know you reviewed and ms that looked at methods and parameter space--can we say anything here?

We have pointed out that the relative performance of these methods is really something of an open question and we encourage people to explore this in the future.

Line 397: “…but infer haplotype blocks only as an incidental output.” In my understanding, the output of Relate tells which branches existed over which genomic intervals and maps specific mutations to specific branches. The shortcoming of this type of output data for identifying haplotype blocks, as defined in this manuscript, is unclear.

Commented [29]: Dasha and Frank: any thoughts on this?

Figure B2 top panel: axis label “Reference Haplotypes” is misspelled

Fixed

Reviewer: 2

Comments to the Author
In this paper, the authors propose a new population genetic definition of “haplotype block” based on the ancestral recombination graph (ARG). The authors outline the idea behind this definition, illustrated examples using schematics, simulations, and empirical data. They discuss the utility of this new definition relative to the ad-hoc definitions used before, and outline how it might be used in the context of currently available tools (e.g. for inferring the ARG).

I’m going to be honest: I struggled to understand a lot of the finer points being made in this paper. However, I would argue that I share a lot of the background and interests of a typical Mol Ecol reader. I work with empirical population genetic data every day, and I am familiar with the ideas behind the ARG, the coalescent, etc. Anyway, in my opinion, if Mol Ecol is now the target venue, this paper would greatly benefit from more basic explanations of many of the theoretical population genetic ideas being discussed here. It seems likely that Mol Ecol was probably not the author's first choice, but I digress. In particular, the figures, while visually attractive, are very hard to understand. I realize that the authors were likely specifically attempting to use these visuals (and the boxes) as an aid to understanding the underlying theory, but I don't think they have been completely successful. Overall, it seems like an important contribution, but I fear it will be largely impenetrable to the merely mortal readership of Mol Ecol. I've tried to outline some ideas for places to address this issue below.

Commented [30]: Would be helpful to identify points of confusion

Commented [31]: Maybe, there could be a box on coalescent theory?

Commented [32]: I think it's better to elaborate on these things in the text--we can also do a better job of point to papers like Magnus's paper on the coalescent in nature genet. (i.e. see xxx for a more in depth description of x).

We sympathize with the conceptual challenges highlighted by this reviewer, because we ultimately give a very different perspective on the ways one might approach empirical sequence datasets in light of new sequencing methods, theory and exciting new analytical tools. Although the manuscript is intended to push the reader to think about things beyond the status quo, we agree that more explanation on some of the fundamental principles will help facilitate understanding of the points that we are trying to make. We have made significant changes to the text throughout to improve clarity and readability. Some of these changes coincide with detailed comments from all 3 reviewers.

Commented [33]: NB: Well, implicitly we argue that existing tools are not enough. A worked example would be great, but how much work is needed??

The Figures

Figure 2 was interesting, but ultimately too complex for me to derive much understanding from. I did not understand the following:
- What is the connection between the colors of the branches and the colors on the DNA sequences. For example, why are the top branches on the tree on the right side of “A” orange, but they lead to red sequences (this is not the case elsewhere?).

This comment reflects an error in the coloring of the branches. This has been fixed.

- What specifically is meant by “light grey branch in both trees shows the effect of recombination in the genealogy” – how does this show the effect of recombination? Recombination between the sequences depicted, or some other hypothetical sequences?

We have clarified this in the captions, but I'm guessing that this will already be much clearer now that the coloring is fixed (see comment above)–this must have been very confusing!

Figure 3 again was interesting but again was difficult for me to understand. Notes below:
- What are the axes in the “SNPs” section? (x and y?)

This has been added to the figure

- “The trees (a - o) show all of the unique topologies that coincide with the genomic spans shown in the central panel (also labeled a- o)”. What is meant by “genomic span” in this case? For example, “O” seems to not be a span, but a single (end) point on a haplotype.

This has been clarified in the caption and main text. O is actually a span, but just a very short one!

- The “L” span appears to contain no SNPs, but has a corresponding tree with internal structure. Is the tree the genealogy (independent of allelic content), or something else?

This has been clarified in the caption, but yes: the tree shows the genealogical history of this region of the genome, but just by chance, no mutations happened to occur in this region.

How the definition would actually work in practice. The authors briefly discuss some of the practical applications of their ideas. I think it would be really useful to actually provide a worked example of how an empirical dataset could have its haplotype blocks, as defined here, and using the existing tools discussed, identified.

Commented [34]: lets see about this...

Reviewer: 3

Comments to the Author
Shipilina et al. provide a new definition for the concept of “haplotype block”, which is based on the ancestral recombination graph (ARG). They suggest that the proposed definition better captures genomic variation than other available definitions and use simulations to highlight this.

I found the manuscript to be fairly well written (although I list a few comments below) and the authors raise several interesting points. The main contribution of this work is the proposed ARG-based definition of haplotype block. I agree with the authors that this definition is meaningful and deserves to be highlighted, although perhaps my main reservation is that this seems a bit limited in scope.

One high-level thought about this work is that working with summaries of the data usually involves a trade-off between how easy these summaries are to compute (for instance, the frequency of an observed allele vs the ARG) and how informative and analytically tractable these summaries are for a particular analysis (see for instance PMID:26395773 for an example of “LD-blocks” that are informative about haplotype structure while remaining fairly easy to compute). The need to work directly with inferred ARGs seems to shift the effort from defining and computing meaningful ARG summaries to the non-trivial task of inferring the ARG itself. It is probably worth further highlighting this trade-off, perhaps in the discussion. Having said that, several methods are being developed to perform accurate ARG inference, so I agree with the authors that the proposed definition is becoming of practical interest. A related comment is that I find the relevance attributed to the proposed definition a bit overstated at times, for instance “haplotypes and haplotype blocks should be the core concepts through which we understand population genetic processes”.

We agree that the trade-off between effort and gains in power to make inferences will play a role in developing practical applications. However, our emphasis is on how to describe and understand the fundamental structure of the ARG, which we believe should be made clear before we think about specific applications. Thus, we stand by the sentence quoted above, that haplotype blocks (or more precisely, observable edges) should be core concepts for understanding population genetic processes.

Commented [35]: I assume her that we have shifted to "observable edges". We could add in the discussion a clearer explanation that in practice, we might work from summary statistics rather than an inferred ARG, and that there is a trade-off here.

I have the following specific suggestions and thoughts:

- I found the paragraph starting at L43 a bit confusing. The authors list several examples of challenging analyses where haplotype structure is important (selection scans, polygenic selection, polygenic scores, gene flow), but the implied connection with haplotype blocks is a bit vague.

Commented [36]: i till need to add something on inference of sweeps...

Commented [37]: All: maybe we could elaborate here and also cite more examples where haplotype info has enhanced our ability to estimate params? Not sure...

- The authors provide a definition of IBD sharing that is based on the existence of a founder population (L131). However, other coalescent-based definitions of IBD exist (reviewed for instance in Wakeley and Wilton, “Coalescent and Models of Identity by Descent”, 2016). The Carmi et al paper cited, for instance, uses this kind of definition.

We have added these references to the text mentioned that more wide use of the IBD approach.

- On a related point, the key element of a haplotype block in the definition provided by the authors is the presence of a coalescence event involving several individuals in a specific region, corresponding to a shared ancestor in the ARG. I suspect an equivalent definition to that proposed by the author may be provided using IBD sharing, as all individuals involved in a haplotype block will be IBD in the region. I don’t think the authors need to try to come up with this kind of definition, but might want to highlight a connection.

Yes there is a connection in that sense, but the way we define blocks would not be the same as a definition based on IBD sharing. But, the key point (illustrated in Fig 2) is that the IBD definition depends on an arbitrary decision about where to cut the genealogy, which subsequently fixes the number of haplotype blocks. The differences with the ARG approach are that (a) the haplotype blocks are determined objectively from the ARG (b) and that ARG blocks can be nested, such that the multiple blocks can be defined based on sub branches of the ARG. This should become clear as the reader moves on and we explore the implications of the definition.

- Fig 2A: I found it confusing that haplotypes for the yellow tree are red.

This was a mistake with the coloring. We have fixed this.

“This branch is defined by a unique coalescent event” it took a few readings of the definition to understand that this coalescent event refers to the lower end of the ARG branch (I assumed it referred to the upper end, which is not unique). Some rephrasing to clarify this earlier on in the definition might help.

We have elaborated on the relationship between coalescence and the branches in the arg. This relates to reviewer 2's request for more explanation about basic population generic processes.

- L183: specify N is the effective size.

Done

- Fig 3: labels of tree leaves are very small.

We have increased the size of these.

- The authors may want to restrict the definition of haplotype block to non-trivial, non-singleton ARG branches (or ARG branches with lower end > 0). Or perhaps not plot them in Figure 3.

We could remove some singleton branches from the figure, but we actually think it will be harder to understand the trees if we remove some of the branches. We have therefore decided to leave this as is. Also, these branches are real features of the data even if they are not especially interesting. In practical inference, one would focus on the more substantial branches that carry multiple SNP.

- L260: the authors observe that deep branches are longer (Supp 1). I believe this is also likely due to re-coalescence events, which are less likely on short branches at the bottom (see difference between SMC and SMC' models by McVean & Cardin and Marjoram & Wall).

Commented [38]: Nick? this makes sense but do we want to mention this?

Commented [39]: Yes - I will look back at the section in SI on this and try to clarify. The reviewer is correct that the deviation from a naive expectation is due to re-coalescence events.

Commented [40]: See comment in main text

We have looked at this question more closely, and expanded the corresponding section in the Supplement ("Inverse relation.."). It is true that re-coalescence events will make deep branches longer than otherwise. However, on re-examining the question, we see that the most important feature is that the distribution of spanned map length ("width") and spanned time ("depth") are both highly random, so that a small fraction of branches carry most of the ancestry ("area"), and therefore, of SNP. We have reworded the text accordingly.

- The discussion section felt a bit rushed and could probably be improved. For instance the paragraph starting at L401 may be moved to the discussion instead.

We have spent some effort to tidy this up.

Commented [41]: Need to be more specific here - once the text is finalised we can go back & do this

- L826: is Li and Stephens used in ARGWeaver?

Commented [42]: Dasha/Frank?

EXTRA:

> Hi Frank, sorry it took me a while to summarize the feedback -- I passed on your comment that you were looking for critical feedback which is reflected in the below. An interesting read and we learned a lot!
> Molly
> Summary of lab feedback:
>
> Major comments
>
> · Throughout the paper, especially in the introduction and conclusion, people felt that there important motivation missing about what knowledge of "true" haplotypes (i.e. their identities in the ARG) or use of a common term for haploblocks would do to advance the field. People did not find the framing of confusion about exactly what different people mean when using the term

haploblock compelling, they wanted it to be related to a biological or technical problem more clearly.

I find this kind of comment worrying, in that it suggests that there is a simple ‘practical’ task (to find selected loci, or estimate Nm, etc); surely we first need to understand what generates patterns in the data (specifically, blocklike haplotype structure), so that we can then try to understand the variety of processes consistent with what we see. Maybe we can emphasise more, at the beginning, that we are aiming at this kind of basic understanding of the relation between the SNP haplotypes and the underlying ancestry, which we need before we can start thinking aboiut what influenced that ancestry.

- Does the use of different terms really belie a misunderstanding of some kind? Is the language leading us astray in our understand of the concepts?

Well, maybe (though I doubt it)each individual is using terms like “haplotype block” consistently - but we found it hard to work out what these different usages might be

- People felt that there was a lost opportunity in connecting the concepts discussed to the literature on admixture (not surprising for our group!) and the literature on balancing selection. There are very clear connections to the ideas of founder populations, inferring the ARG, and many of the same problems arise (e.g. “invisible” recombination events between the same ancestry states). We felt that the term ancestry tracts perhaps has less confusion around it and has nice connections here to what is meant empirically by haplotype.

There is a connection, in that the “naive” IBD definition that we start with is defined by ancestry in a reference population. However, I find the world of admixture models (STRUCTURE etc) quite confused: one gets the impression that people believe that present-day populations really are made up of a mosaic of real “ancestry tracts. It seems to me that this is hardly every the case, even in hybrid zones - instead there is an accumulation of successive admixtures, which may not be possible to disentangled. Maybe we should say this, and contrast with our view which is based on “branches” - which are real sets of ancestral sequence.

We also thought the literature on balancing selection would be a nice connection to ideas about the age of tracts, the timing of their coalescence, and the density of informative sites.

Maybe a nice example, but it seems to me to be adding a complication to an already complicated paper.

- There is also a connection with the literature on ancestry tracts as recombination clocks (e.g. Moorjani et al)

Also Ungereer et al, PNAS 1998

- People had trouble with many of the figures and did not find them intuitive.

Some specific comments:

- Is figure 1B useful? Inversions are not mentioned elsewhere in the paper nor the fact that they trap haplotypes by blocking recombination

§ In general people thought there could be better integration with the type of empirical data people collect and potential shortfalls of e.g. statistical phasing, Fst, etc

We can mention this more explicitly, but we are dealing with more fundamental questions & shouldn't get bogged down with technicalities

○ Figure 2 – we got confused about what was being shown – the recombination event depicted didn't look possible to us with the haplotypes shown. How does what is being shown here interact with the timepoint argument?

○ Lots of confusion surrounding Figure 3 and what was meant by shallow and deep, though we were able to work through it as a group. The flow from the trees through the blocks was subtle and the indication of time in the block component of the figure also difficult for people to follow. There was some duplications in the letter labelling and in general people felt there was too much going on to follow it clearly

§ Many people wanted this figure to be simplified and have fewer panels

I think that it's important to have a complete figure, showing all the genealogies - which makes it possible to work through and undertsasnd what is happening.   Perhaps we could say this explicitly.

· People felt that there could be improved clarity throughout the manuscript about when empirically inferred versus theoretical concepts are being referred to (e.g. lines 158 in terms of the definition of IBD). People were unclear on whether you were trying to make a distinction between haplotype blocks, IBD, and coalescence throughout the manuscript or whether they were intended to be synonyms in the sense that they link to the same processes.

We do need to clarify this! Each of these terms does have its own definition, but they are not the same at all.

· People were unsatisfied with the conclusions section. They felt that this section was light on solutions/ideas in that is mainly suggested new methods and analysis. They wanted more of a balance of problems and solutions. In particular several people wanted to know what the big unanswered questions are that we could tackle if we knew haploblocks with certainty, how would this change our thinking and the field.

I think everyone agreed that haploblocks as described were a fundamental unit of evolution but what are the big advances that could be made if we solved these technical/conceptual problems.

Actually, it's not obvious to me how much we can gain from knowing the full ARG. One way to frame this would be to say that under a simple model (eg a hard sweep) there is a (fairly) clear prediction for the pattern that we expect to see. So, knowing the full ARG would give more accurate estimates if this simple model is valid. More likely, though, is that we would find that it is not - eg because there were multiple overlapping sweeps. It is quite unclear how far we will be able to disentangle such scenarios.

- One person mentioned testing qs about hybrid speciation as a potential application

<u>Minor comments:</u>

-ARG controls for population structure better – instead of solving problems with GWAS – can be used to better correct for populations structure (relevant papers are cited later)

-Difficulties in understanding caused by inconsistent usage of abbreviations

-Some found banded blocks of SNPs confusing as a framing

-Figure B1 was helpful

-Figure 4 – says something about the simulations being based on haploids, people wondered if you mean haplotypes

It doesn;t make much difference that simulations are of haploids but we need to say this

-What are the consequences of the assumptions in the LS model? How does this lead us astray?

Probably not much?