Response to the editor

Thank you for considering this manuscript for publication. We are glad that Reviewers 1 and 2 are enthusiastic about this community effort, and their reviews have helped to resolve a number of issues with the manuscript. We are disappointed that although Reviewer 3 agrees in the value of the software, they do not believe the manuscript is appropriate for publication in Genetics. However, as we point out in the response, it is unclear whether they were reviewing according to the "Methods, technology, and resources" section scope.

Specific comments

It is most important that you address the following in your resubmission: Improve the structure and focus of the main body of the manuscript. This should include moving a significant portion of the material to appendices or supplemental material, and adding an up-front summary of the main new features of msprime (relative to the 2016 paper).

We have shortened the manuscript by moving the details of the analysis of Hudson's algorithm and selective sweeps to appendices, and deleting unnecessary text in several sections. We have added Table 1, which summarises the features added to msprime since the 2016 paper.

Response to the reviewers

We thank the reviewers for their close reading of our manuscript and insightful comments. In the following we address the points raised in turn.

Reviewer 1

Reviewer Point 1.1 — Simulation is a vital part of population genetics, and msprime has quickly become a crucial tool for many researchers. msprime was first presented in a publication in 2016, where the main focus was the use of "sparse trees" and "coalescence records" to greatly increase the efficiency and practicality of large coalescent simulations. In this new manuscript, the authors introduce a new version of the software, with this manuscript ""focusing on the aspects that are either new for this version, or in which our approach differs significantly from classical methods."

Since 2016, msprime has gained popularity not only due to its computational efficiency, but also for its software development style with well-documented features, development guidelines, and ever-growing capabilities. The extensive author list shows how the community around msprime and associated programs has grown over the last five years. The current manuscript has a huge number of citations (I count 178!), generally representing a willingness to engage with a broad spectrum of the adjacent scientific community.

The manuscript covers many features of the 1.0 version of msprime including at least (my list and organization):

Software / API / innovations: Clear separations of ancestry and mutation simulations Improved methods for downstream analyses Integration with other software including other simulators Estimating run times of sim_ancestry ARG recording Specific API for specifying demographic histories Software development model

Implementations of external methods: Selective sweeps Instantaneous bottlenecks Multiple merger coalescents Discrete time Wright-Fisher Gene conversion Mutation models

As well as benchmarks against other software implementations.

The manuscript is well written and comprehensive - I do not find any major deficiencies. This manuscript handles a wide range of topics with a high degree of clarity. And I find the manuscript certainly reaches the criteria "... tools of interest to a wide range of geneticists" and so is a good fit as an investigation within the "Methods, Technology, & Resources" subject.

Reply: Thank you for the comprehensive summary. We are delighted that these community efforts are so well received.

Reviewer Point 1.2 — However, I do think it would be possible to increase the clarity and usefulness of the manuscript with some relatively small changes in organization.

Reply: Thank you for the detailed review and suggestions, which we hope have resulted in an improved manuscript.

Reviewer Point 1.3 — As, is, the manuscript lacks a simple overview of the new capabilities of version 1.0. Instead these are described in separate sections. The manuscript might benefit from a short summary and or table of the software's basic capabilities, API innovations, new functionality, etc. This may also provide a way to structure the larger flow of the paper, which currently seems to touch on the different topics in a somewhat random order.

Reply: We have added table 1 to summarise the new features.

Reviewer Point 1.4 — Generally, I found the figures and examples appropriate and well done. However, I think they could benefit from more cohesive presentations and descriptions. I will provide a number of examples - figure 3 and figure 4 present running times of various aspects of msprime, but present similar data in different ways. In Figure 3, the data points are not presented directly, while they are present on Figure 4. And based on the descriptions sometimes replicates are summed over (Figure 7) or averaged over (Figure 8?) to determine the values on the y-axis. Figure 3 clearly specifies sim_mutations as the input for the benchmark, while for most other figures (e.g. 4) it frequently just says something like "running time for msprime." Sometimes only population-size scaled rates are provided in descriptions (e.g Figure 5 and 7) while in others both scaled and absolute sizes are provided. Sometimes "samples" are described to be diploid (figure 1) or sometime haploid (Figure 3), but other times the ploidy is left implied (e.g Figure 4). Sometimes the population size is referred to "effective population size" (e.g Figure 4) and other times just "population size" (Figure 3). More consistency in these simulations, descriptions, and figures could add clarity to the presentation.

Reply: Thank you for examining the figures in this detail and providing such helpful feedback. We have tried to improve the presentation by taking the following steps for the main text figures:

- All plots now contain data markers to help show where the data points are and to distinguish the lines from each other.
- All plots state they are running "sim_ancestry" or "sim_mutations".

- Stated explicitly whether samples are haploid or diploid in the figure captions, and updated axis labels also.
- All figures now show the time averaged across replicates.

In addition figure 4 has been moved to the appendix, making the remaining figures much more uniform in appearance and purpose (it is a special case because it is showing how well theoretical predictions map to observations, as well as those observations).

Reviewer Point 1.5 — I did find the lack of a disk storage benchmark an interesting choice, as this aspect of the software is highlighted in the introduction but is not present in the manuscript.

Reply: We did not include any benchmarks of storage space as this was explored in detail in the 2019 Nature Genetics paper, and we directly reference this paper when discussing the storage efficiency of tree sequences.

Reviewer Point 1.6 — Figure comments Figure 2 I think it could be made more clear that the genotype matrix is not part of the set of tables.

Reply: We agree, and have revised Figure 2 and the caption to clarify the distinction between stored and derived data.

Reviewer Point 1.7 — Figure 4 It is not clear if these two plots present all the data used for the quadratic fits, or if these data continue outside of the plots. Samples are described, but not their ploidy, are they haploids or diploid? From the plots here, it is not clear if the quadratic relationship holds for sample sizes $\ll 1000$. I suggest this figure should also include a smaller sample size, (eg n=10)

Reply: We have moved Figure 4 to the appendix, as it is of more specialised interest than the other main text figures. We have truncated the x-axis on the quadratic fits to just the available data, to clarify that we are showing all of the data. We have clarified that samples are diploid.

The plot is already quite complex and it is difficult to see how we could add a third line without major changes. The figure is primarily intended as a useful yardstick for users who wish to run simulations and would like a rough idea of how long it should take, and therefore adding a third line for very small sample sizes is likely to make it less useful for this purpose.

Reviewer Point 1.8 — Figure 5 I suggest including the absolute Ne and gc rates be included in the legend, in addition to the scaled rate.

Reply: We have redone this figure with the exact estimates from Lapierre et al and quoted the absolute Ne and gc values.

Reviewer Point 1.9 — Figure 7 Why is an order of magnitude lower recombination rate used here than in other examples? This choice seems arbitrary and makes it difficult to relate these benchmarks results to any others in the paper. Samples are diploid or haploid? I would suggest the absolute Ne and s rates be included in the legend, in addition to the scaled rate.

Reply: We have redone this figure to use the same recombination rate and population sizes as the rest of the paper, clarified that samples are diploid, given results in terms of the average per replicate and generally made the plot as close to the others as possible.

Reviewer Point 1.10 — Figure 8 Averaged points are referenced but not shown "Each point...". The plot could be more clear where the data points are and where the lines are interpolating between data points. Because of how the simulations were described and plotted, I found it awkward to try to compare the running time of the DTWF and coalescent models, even if this was not the purpose of the current figure. The figure says "to ensure we are measuring" I would suggest changing the text to "with a goal of measuring".

Reply: We have update this figure (and all others in the main text) to include data markers. We have also rephrased the caption to try to clarify the parameters simulated and make them consistent with other figures.

Reviewer Point 1.11 — Individual line comments Line 142 - what is the "msp" program? Is this a command line version?

Reply: We have clarified this point with a parenthetical comment.

Reviewer Point 1.12 — Line 370 - the text here makes it sound like Kelleher et al (2016) makes a statement relating eq. 1, but the Hein et al (2004) is not cited in the Keleher et al (2016).

Reply: We have rephrased this to say "(see also Kelleher et al., 2016, Fig. 2)".

Reviewer Point 1.13 — Line 370 - population size(s)

Reply: Thank you, we have corrected this sentence in the revised manuscript.

Reviewer Point 1.14 — Line 420 - Despite the long description, I was not able to follow how the previous paragraph implies that "work is spread out relatively evenly on the coalescent time scale"

Reply: This is implied by the fact that the number of lineages decreases slowly on the coalescent timescale. However, the point is not crucial to the main goal of the paragraph, and we have therefore deleted the sentence for clarity.

Reviewer Point 1.15 — Line 422 - There is no context to what is meant by "large" or "small" here - is 2 or 10 or 10000 "large".

Reply: TODO

Reviewer Point 1.16 — Line 607 - The equations for the allele frequencies during sweeps seem slightly out of place as (I assume) these are not novel to msprime. There are many places where more details could be provided about methods discussed here, but in other cases the equations are left for the supplement or other papers.

Reply: We have moved the sweep trajectory equations to an appendix.

Reviewer 2

Reviewer Point 2.1 — Kelleher and colleagues present the 1.0 version of 'msprime', the leading coalescent simulator which is built upon a highly memory- and time-efficient data structure, introduced by the lead author, which supports exact simulation under the coalescent model with recombination.

In addition to being highly efficient, the software engineering of msprime is of the highest quality. The program includes extensive unit and validation tests, and offers an API that simplifies worksflows and reduces the inefficiency and sources of error associated with large intermediate text files.

Kelleher has made every effort to engage with the community, with the result that msprime now support many specialized features that have been contributed by other developers. This has the important benefit of reducing the chances that this software will lack support in future.

In short, this is an admirable project, and a shining example of successful academic software development.

Reply: Thank you for the kind words, we are delighted that these community efforts are so well received.

Reviewer Point 2.2 — The paper itself however is less polished. It is very long (37 pages / 1200 lines without appendix), mostly because it is in places overly detailed, and seems to err on the side of completeness rather than conciseness or clarity.

Reply: Thank you for this well-founded criticism. We have reduced the length of the manuscript by moving some material to appendices, cutting "documentation-like" text, and tried to address specific points raised below. We hope that this has resulted in an improved manuscript.

Reviewer Point 2.3 — For instance, lines 160-177 describe the data structure in some detail, while this is published information. Line 192 remarks that "storage space is dramatically reduced", and goes on to make this more precise and point out some (fairly obvious) advantages in lines 193-202.

Reply: It is true that the details of the succinct tree sequence have been laid out in earlier publications, and it is reassuring that the advantages are obvious to you. However, for many of the intended audience these methods will be new, and we feel it's important that this paper is a self-contained, thorough, and convincing argument for the advantages of tree sequences and the tskit library in simulation workflows. Very many papers are still published using ms and other simulators that use inefficient data formats, and it is these users that we want to convince. We also hope to convince those that might be developing their own simulators to take advantage of tskit and its extensive functionality, rather then developing everything from scratch, as is the classical and still dominant approach.

Reviewer Point 2.4 — As another example, line 301-4 states that "Simulating mutations ... is efficient" and refers to Fig 3 for evidence. The rest of the section (304-319) details a number of examples shown in Fig 3 that do not add much to the story.

Reply: We have removed the some of the examples and shortened the section.

Reviewer Point 2.5 — As a third example, in the recombination section after useful comparisons with other approaches, line 354 states that the proposed algorithm is still quadratic in the recombination rate. The long section from line 354 to 426 details the algorithmic reasons, and use various analytic approximations and simulations to support this. This material is of interest only to a few algorithm developers, and I would imagine it would be better placed in an appendix, with the key observations summarized in the main text.

Reply: We have moved this detailed discussion to an appendix.

Reviewer Point 2.6 — Another feature of the current paper is that a number of topics return several times with minor variations. This is particularly true for algorithmic efficiency, and the API, both of which are discussed multiple times. It is of course true that algorithmic efficiency is the major distinguishing feature of msprime; nevertheless, the focus of the current paper seems to be the rich feature set, and the software development model, as well as the various APIs that allow for a more integrated and less error prone workflow. I would suggest to introduce separate sections for algorithmic efficiency and the API, and discuss the various features (as you do now) in their own separate section, but focus on these features from a user's rather than a developer's perspective.

Reply: The large number and diversity of features in msprime has made the organisation of this manuscript a major challenge, and the present layout of a series of relatively self-contained topics arrived at after much experimentation. Issues like interface design and algorithmic efficiency necessarily cut across most of these topics. Separate sections on efficiency and API design would need to refer to these other sections for context, and would lead, we believe, to a longer and less cohesive document.

However, we do agree that some sections are too long and this criticism has been very helpful to allow us see where content could be cut without loss of information. For example, we have cut out references to the API design in the Simulating Mutations section, and substantially cut back the Demography section (see next point).

Reviewer Point 2.7 — Occasionally the paper reads more like documentation than a paper, such as when the Demography class is introduced and several associated functions are mentioned - this is not relevant for a paper in my opinion.

Reply: This is a very helpful observation, and helped us clarify where text can be cut. We have substantially cut back the Demography section and other areas where we were veering towards documentation.

Reviewer Point 2.8 — The section about the development model is interesting - very few papers include such a section, and indeed I think this is one of the stand-out features of this project that it is so successfully supported by many contributors. My only suggestion here is to remove the listing of number of lines of code - this is a very questionable proxy for quality or quantity of work, and the work is impressive as it is.

Reply: Thank you for your suggestion, we have removed the number of lines of code in the revised manuscript.

Reviewer 3

Reviewer Point 3.1 — This manuscript introduces a simulation tool, msprime version 1.0, and describes features, additions, and performance of the simulator relative to the earlier msprime software. I will begin by saying that I believe all researchers in population genomics who use simulations have appreciated the quality work that has been previously done in the msprime sphere. The contributions proposed here clearly improve on a number of features, and that of course has value.

Reply: Thank you, we appreciate the kind words.

Reviewer Point 3.2 — However, I did not find this work to be appropriate for Genetics, and would rather view it as more appropriate for a journal like Bioinformatics. The fundamental reason is that this basically appears to perform old functionalities faster, and the added functionalities are modest. Again, both are valuable, but this work doesn't address much larger biological / statistical inference problems that are in need of development in this area, which could justify publication in a journal like Genetics.

Reply: We respectfully disagree, and wonder if there is perhaps a simple misunderstanding. While we absolutely agree that if this manuscript were submitted as an "Investigation" to GENETICS, it would need to make substantial methodological advances of the type that you describe. However, we are submitting the manuscript for consideration in the "Methods, Technology, and Resources" section, which has the following description:

> GENETICS welcomes manuscripts that describe genetic or analytic methods or resources that are likely to have broad impact. They can be full length research articles or Communications. The method or resource needs to be novel, or be a significant advance in an existing method or resource. It should be of considerable interest to a wide range of geneticists or of extraordinary interest to a smaller group of geneticists. The method, resource or technology should enable experiments that will allow investigators to address significant biological questions and should be described well enough so others can implement the method. The necessary reagents or resources need to be available upon request.

We believe that this manuscript amply meets these requirements:

- The software is likely to have broad impact—the 2016 paper has been cited 123 times so far in 2021 according to Google Scholar.
- This is a significant advance over the 2016 paper, with a great deal of new functionality (see the newly-added Table 1).
- It is of interest to population geneticists who use simulations.
- The functionality enables novel simulation types (e.g., Λ-coalescents) and greatly increases the scale over which existing methods can be applied (e.g., bacterial simulations).

Reviewer Point 3.3 — Specifically, the need of the field is really to simulate more realistic models in a coalescent framework, rather than the slower forward-in-time choice that stands as the only option presently. For example, the practical use of a tool that only simulates strictly neutral histories, or histories with the addition of positive selection, is likely rather limited empirically. By which I mean, the functional regions of a genome of course experience purifying selection , but there is growing appreciation of the fact that this may result in background selection effects that may be widespread across the genomes of many species. In order for the strictly neutral demographic simulations to be of value, one must identify genomic regions that are not only neutral but are also unaffected by selection at linked sites. In a great many organisms, such regions may well not exist at all; in terms of commonly studied species, these regions may only exist in a handful of large, coding-sparse vertebrate and plant genomes (though, even if these regions exist, they are not necessarily easy to identify, or sufficiently large to perform neutral demographic estimation). I note that the manuscript avoids any mention or discussion of this problem, and avoids any citations that quantify this issue, but the neglect doesn't alleviate the problem. The authors should really consider, discuss, and utilize the following work carefully in this regard:

Lohmueller et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLOS Genetics.

Comeron. 2014. Background selection as baseline for nucleotide variation across the Drosophila genome. PLOS Genetics.

Elyashiv et al. 2016. A genomic map of the effects of linked selection in Drosophila. PLOS Genetics.

Comeron. 2017. Background selection as a null hypothesis in population genomics: insights and challenges from Drosophila studies. PTRSB.

Pouyet et al. 2018. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inference. eLife.

Torres et al. 2018. Human demographic history has amplified the effects of background selection across the genome. PLOS Genetics.

Campos & Charlesworth. 2019. The effects on neutral variability of recurrent selective sweeps and background selection. Genetics.

Jensen et al. 2019. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution.

Castellano et al. 2020. Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other Homininae. GBE.

Torres et al. 2020. The temporal dynamics of background selection in nonequilibrium populations. Genetics.

Johri et al. 2021. The impact of purifying and background selection on the inference of population history: problems and prospects. MBE.

Murphy et al. 2021. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. bioRxiv.

Reply: We agree that neutral simulations are an approximation, and care needs to be exercised in their use. We have added a sentence to introduction to emphasise this point. However, as we have argued above, major methodological advances of the type that you discuss (which we agree are extremely important!) are beyond the scope of the current manuscript.

Reviewer Point 3.4 — Relatedly, for the simulations of positive selection, while it has certainly

been common practice in the field for a few decades, it is rather odd to model a single positively selected site on a background of all other neutral sites. Positive selection may act on mutations in functional regions, and functional regions will be largely shaped by purifying selection. Furthermore, the Hill-Robertson effects between the modeled positively selected mutation and the neglected array of negatively selected mutations is often non-trivial. In that sense, the positive selection model is probably only of empirical relevance in organisms with something approaching free recombination (to eliminate such linkage effects), as in perhaps HIV or the like. But of course, these sorts of organisms tend to have small, coding-dense genomes and experience very strong selective pressures, thus returning to the first problem mentioned above that regions free of background selection effects may not exist in the first place.

Reply: We agree that this is an important point, but again this is a deep methodological issue that is beyond the scope of the current manuscript. We have simply implemented methodology that has been used and studied decades. Even if the goal is to demonstrate that such approximations are deeply flawed, it is still necessary to be able to simulate them efficiently and conveniently.

Reviewer Point 3.5 — In summary, I believe this to be a nice software addition that belongs as a Note in a bioinformatics journal. The software solution that would represent an important biological advance for Genetics would involve simulating more realistic models as discussed above (as forward simulators are currently capable of doing), but at a fast enough rate to be applicable to large-scale simulation-based genomic inference.

Reply: Thank you for the detailed and interesting review—we agree that there is much work to be done in the field of population genetic simulation! As we have argued above, we believe the manuscript is entirely suitable for publication in Genetics as a "Methods, Technology, and Resources" article.