Author response

We thank the editor and the reviewers for helpful comments. Below is our detailed *response*. To summarize the main changes, we have

1. revised the entire manuscript with the Molecular Ecology readership in mind and to emphasize the biological interpretation of the models and the results as requested by the editor and reviewer # 1.

2. moved old Figures 1, 3, 4, 6, 8, 9 to Supporting Information and deleted old Figure 7 to reduce the number of figures in the main text as requested by the editor

3. moved old Eqs 1, 5, 6, 7, 9, 10, 11 to Supporting Information and the equation between old lines 200 and 201 to reduce the number of equations in the main text as requested by the editor

4. added Fig 1, a simple illustration of the difference between random sweepstakes and the Wright-Fisher model, as requested by the editor

5. added Fig 2, a simple illustration of how random sweepstakes can speed up fixation given that fixation occurs, as requested by the editor

6. added legends to the graphs to explain the relevant parameter values for each panel

7. Furthermore, we now explain on line 418 and in each relevant caption that the trajectories in each panel were obtained under identical conditions

8. The manuscript on the Atlantic cod data analysis (Árnason et al., 2022) became available on bioRxiv on June 16, 2022; the manuscripts by Chetwyn-Diggle et al and by Dahmer and Eldon are still not available so we have removed references to them.

Response to Editor

1. We have now had three reviews from experts in the field. All three were very complementary of your expertise in the field and your knowledge of the topic, which is highly relevant to the current issue. However, all of the reviewers, which included theoreticians, thought the manuscript was dense and difficult to understand for the average reader of Molecular Ecology. Reviewer 1 thought the paper would be better suited for a mathematical biology journal and that it would require too many substantial changes to be suitable for Molecular Ecology. However, if you can transform the manuscript into work that has broader appeal, I would welcome your resubmission. A paper that translates some of this theory into a form that is more easily digestible may be well cited and aid the dissemination of these ideas to a broader audience. For example, the explanation of the effect of sweepstakes reproduction on rapid adaptation would benefit from a figure or box illustrating using a cartoon how this mode of reproduction is different from the standard fisher-wright model and how it impacts rapid adaptation. Two of the reviewers pointed out that you rely on unpublished work that is not even available on a pre-print server, and this needs to be addressed

and the methods more clearly articulated.

Author response: Thank you for the opportunity to revise the manuscript. We have revised the manuscript with a broad audience in mind. Specifically we have added Figure 1 to explain the idea of random sweepstakes and how it is different from the Wright-Fisher model, and Figure 2 on how random sweepstakes speed up the time to fixation conditional on fixation. The manuscript (Árnason et al., 2022) became available on bioRxiv on June 16, 2022 (https://doi.org/10.1101/2022.05.29.493887); the manuscripts by Chetwyn-Diggle et al and Dahmer & Eldon are still not available as preprints and so we have removed them at the recommendation of Reviewer # 2 and # 3.

2. I also strongly recommend that you reduce the number of equations and figures in the text. Please consider placing some in them in the supplementary. The figures could be vastly improved. First, there are too many (10 Figures). Second, the parameters are not explained in the figure legend. Third, what do all the different line types mean? In Fig 7 what is b? Why not be more explicit and explain what "eq 4" is instead of constantly referring to it. There are also frequent short forms in the legends without explanation (e.g., "resp."). The figure legends should explain what is going on in the figures without the need to refer to the text. Figure 10 is not well explained. Finally, there are many different panels in most of the figures, and it isn't clear without careful dissection of the legend what the differences are. This could be fixed by reducing the panel number and labelling the panels to clearly identify what the point of the different panels is.

   Author response: We have expanded on the explanation of Eq (4); as detailed above we have moved many of the equations and figures to supplementary material. We have removed old Figure (7) as it overlaps with Figure S5. We have improved the figure legends. We now explain on line 418 (as we begin to show the results), and in each relevant caption, that the different line types and colours are only meant to distinguish between the trajectories; in each panel the trajectories were obtained under identical conditions. Furthermore, we now extensively explain the model parameters.

Response to Reviewer # 1

1. Firstly, the presentation is far too technical to be accessible to a typical Molecular Ecology reader. Even the most important results of the model, which could be presented very clearly and in plain language, are often buried in a mass of details that obscure the main messages.

   Author response: We have revised the entire manuscript to make it more accessible to the Molecular Ecology readership; for example starting at line 176 we have added a biological interpretation of the models of random sweepstakes as in Eq (3). In a simple illustration we now explain the difference between the Wright-Fisher model and random sweepstakes, see Fig 1 by line 659. Between lines633 and 638 in § 5 we briefly summarize the main findings from the simulations.

2. Secondly, I don't think the authors have made a compelling case that there are serious problems with the standard theory (as is suggested towards the end of the paper). At best, the authors have simply raised the possibility

that their models could be important and deserve further attention by theoreticians in order to fully explore the theoretical consequences of sweepstakes reproduction.

Author response: We believe that the results of the analysis of the Atlantic cod data(Árnason et al., 2022) $https://doi.org/10.1101/2022.05.29.493887$), which we review in § 4, presents evidence for the conclusion that the 'standard theory' (understood here as represented by the Wright-Fisher model and the Kingman coalescent) is inadequate in explaining genetic diversity in highly fecund populations. In addition, U-shaped site-frequency spectra, predicted by multiple-merger coalescents (and not by the Kingman-coalescent even taking population size changes or background selection into account), is observed across domains of life (Freund et al., 2022). To emphasize this we added the new manuscript of (Freund et al., 2022) $https://www.biorxiv.org/content/10.1101/2022.04.12.488084v2.full$ at line 654. In addition, in the Introduction, starting from line 88, we outline the biological reasons for why the Wright-Fisher model may be a poor choice for modeling highly fecund populations.

3. L 27: as I understand it, the evidence for adaptation for human height (among European populations) is not as strong as once thought. See e.g., Berg et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. Elife 8:e39725; Sohail et al. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife 8
   Author response: The critical analyses of (Berg et al., 2019; Sohail et al., 2019) are now mentioned on line 39

4. L 47: "a shift of the fitness optimum may also be caused by a new mutation". I do not understand what is meant here. Mutations cause shifts in optima? It's not obvious why this should be so.
   Author respons: In the classical hitchhiking model the occurrence of a new favorable mutation is assumed. To avoid confusion, this part of the manuscript at line 59 is deleted.

5. I imagine many evolutionary biologists would disagree with the definition of rapid adaptation as outlined here. Rapid adaptation is said to mean that $2Ns > 1$, yet for large populations sweeps with very small s will meet this condition. The definition of rapid in such cases would be quite slow to many.
   Author response: On line 66 we state that $2Ns$ must be at least one hundred (not one!) to observe rapid adaptation/sweeps.

6. L62-74: I'm not sure how much the tangent into formalism helps in the introduction. But if you must, things would be easier if you use common notation that will allow more people to see that these are fairly standard theoretical results that they have seen before. As far as I can tell, C_2 is the additive genetic variance, which would typically be G or V_A. And $s = 1/(V_P + V_S)$, or roughly $1/V_S$ under weak stabilizing selection, where V_S is the width of a Gaussian fitness function and V_P is the phenotypic variance. Since s has already been used to represent a selection coefficient, why not gamma, which is often used in its place? Also, the basic prediction

(exponential decline of the distance to the optimum, over time) is a far older result than implied here (e.g., you can find it in Lande's 1976 Evolution paper), though I agree that explicit population genetic models of the phenomenon or more recent.

Author response: The paragraph beginning at line 68 has been rewritten without using formulas. (Lande, 1976) is now cited.

7. L84: Here is a major question to motivate the paper and future work, though it is sort of hiding at the end of a paragraph that is pages into the introduction. A statement like this is something that could be highlighted in the abstract and then the authors could also give the main punchlines there (or at least what is currently known). The abstract does mention that sweepstakes reproduction "may facilitate rapid adaptation", but it would be nice to specifically see when it should. You could be more specific about this in the abstract. What are the conditions under which it facilitates adaptation? Are these conditions likely to occur, or do we currently lack the key data that would allow us to know one way or the other?

   Author response: The paragraph beginning at line 88 has been rewritten without giving the main punchlines suggested by the reviewer. Our goal in the Introduction was to show how this project was extended from classical population genetic models of genetic drift and selection to sweepstakes and selection.

8. The paragraph beginning L 175 is a good example of text that will almost certainly be inaccessible to most readers of Molecular Ecology. Equations are fine but they must be unpacked and the biology must be explained in reference to the model. There is little unpacking in this section (and most that follow) which severely limits what readers will be able to take from the paper.

   Author response: We have revised § 2 and added clarifications to explain the biological meaning of the equations. In particular, we have expanded on the explanation of Eq (1) and how it relates to random sweepstakes, so the biological interpretation of Eq (1) and the parameter α should now be clear.

9. L303-307. The wording is confusing here. I believe the point is as follows: IF random sweepstakes reproduction were to facilitate rapid adaptation, THEN we would need it to increase the number of beneficial mutations that enter the population in order to offset the decline in fixation probability that occurs per beneficial mutation under sweepstakes reproduction. Is this the point or perhaps I have missed it? If this is the point, it's unclear if, when or why random sweepstakes would increase the input of new beneficial variants into the population.

   Author response: Yes, this part may have been unclear and we apologize for the confusion; thank you for pointing this out. We are not saying that random sweepstakes are increasing the input of beneficial variants, but we agree that it appears random sweepstakes reduce the probability of fixation. However, our mixture model in Eq (4) of random sweepstakes may moderate the reduction of the fixation probability (relative to Eq (3) alone). We have revised the corresponding part between line 388 and line 398.

10. As a way to conceptualize the issue, perhaps it would be worth clarifying what exactly is a reasonable model for the rate of adaptation. Most of the focus of the paper is on the sojourn time of a segregating beneficial allele, conditional on its fixation. Sweepstakes reproduction can shorten the fixation time, yet it also reduces the probability of fixation to begin with. One could come up with a very different model for the rate of adaptation that would presumably lead to the conclusion that sweepstakes reproduction decreases the rate of adaptation. We could, for example, model the rate of adaptation as the product of input of new beneficial mutations, per generation, and fixation probability of each, as in many models developed since the 1970s (i.e., "origin-fixation" models; McCandlish and Stoltzfus 2014 Quarterly Rev Biol.). In the standard version of the theory, and assuming $1/2N \ll s \ll 1$, where s is the fitness benefit associated with each beneficial allele an individual inherits, the mutation rate per generation is $2Nu$ and the fixation probability is $2s$, making the rate of adaptation $4Nus$. Assuming that $2Nu$ does not systematically differ between Wright-Fisher and sweepstakes populations, then the main difference would be the fixation probabilities in each type of population, and so the sweepstakes population would adapt more slowly. I'm not arguing that this type of model is the best model for the rate of adaptation (see, for example, Gillespie's take in "Why $k = 4Nus$ is silly"). Rather, I'm simply trying to highlight the ambiguity in exactly how rates of adaptation are and should be conceptualized. The authors should lay out a clear case for how they measure the rate of adaptation, and why it is a reasonable way to do so.

   Author response: Thank you for the thoughts. We agree that the phrase 'rate of adaptation' is ambiguous as there are (at least) two quantities to consider as the reviewer points out, the probability of fixation ($p_N(1)$) Eq (S10) and the time to fixation conditional on fixation ($\tau_N(1)$) Eq (S10). We suggest that considering $p_N(1)$ and $\tau_N(1)$ jointly would give a more complete picture of the effect of random sweepstakes on positive selection, and by implication on adaptation. In the Durrett-Schweinsberg model of recurrent sweeps(Durrett and Schweinsberg, 2005a), the selective advantage $s \in (0,1)$ of a new advantageous mutation is independent of the population size, and the probability of fixation of the mutation is approximately s, so is not necessarily $\ll 1$. Note also that here, as in the Durrett-Schweinsberg model, we are concerned with one mutation at a time. As we state on line 59 the rate of response to selection, or the rate of adaptation, is given by the inverse of the time to fixation given that fixation occurs; in that sense with random sweepstakes on average reducing the conditional expected time to fixation we come to the conclusion that random sweepstakes facilitates the response to selection.

11. L327+: Here, where the first major results are presented, the language is decidedly non-biological. Results are presented in terms of specific parameter values without reminding readers what biological scenarios the values correspond to (this is also true of the figure legends, which are hard going). I believe $\varepsilon_N = 0$ corresponds to a scenario where sweepstakes are absent, and if so, this is worth highlighting so that it is obvious what the point of contrast is. If I have misinterpreted this basic result,

then all the more reason to help readers along. The same comments apply to subsequent results and their associated figures, which again emphasize parameters and mathematical notation rather than biological messages.

Author response: We have revised the presentation of the results to emphasize the biological interpretation. Note that we clearly have to state under what conditions (i.e. parameter values) the results are obtained; and the parameter values chosen should only be understood as giving an example of the effect of random sweepstakes and recurrent bottlenecks on selection. In paragraph beginning at line 510 we summarize the simulation results for haploid populations, and in the paragraph starting at line 556 for diploid populations.

12. L356: B has already been used to in the Brownian motion term in eq. (1)
Author response: We have moved this equation (now Eq (S3)) to Supplementary material at the request of the editor to reduce the number of equations in the main text, and now use W_t to denote Brownian motion.

13. L497: "fix certain important problems" Like what exactly? It is not obvious from the paper how problematic standard models are. This is not simply an issue of noting that models make simplifying assumptions. You would also have to make a clear argument for why their assumptions are problematic in a way that the new sweepstakes models are not. It would also have to be clear how much of a difference it makes and that most species are likely to have attributes that make the standard models (whatever they are) inappropriate for describing their evolution.
Author response: As we explain in point # 2 and line 595, the 'standard theory', understood as representing either the Wright-Fisher model and the Kingman coalescent, or the Schweinsberg model(Schweinsberg, 2003) and the Beta$(2-\alpha, \alpha)$-coalescent, does not explain the population genomic data of Atlantic cod at all, see (Árnason et al., 2022). However, we agree that the phrase 'certain problems' is unclear. The problems we are referring to relate to the assumption of the Schweinsberg model(Schweinsberg, 2003) of unbounded fecundity, and the coalescent timescale of the Schweinsberg model; these problems are discussed at line 243 and line 246. We have now revised the text at line 629 and explicitly refer to the two problems mentioned. In this context we are not concerned with 'most species' but with highly fecund populations as stated in the Abstract and on line 90. However, U-shaped site-frequency spectra as predicted by multiple-merger coalescents are observed across domains of life(Freund et al., 2022), as we now state on line 654.

14. L502: Yes, but fixation probabilities also decline, so can you really conclude that sweepstakes facilitate rapid adaptation? Increased failure to adapt is another outcome of sweepstakes.
Author response: see our response to Comment #10.

15. L506+: showing that your model is consistent with genomic data does not seem to prove that sweepstakes reproduction is the cause of rapid adaptation. Or maybe I am missing some step in the chain of logic.
Author response: We are not saying that sweepstakes reproduction causes adaptation, but that sweepstakes reproduction, in this case involving strong positive selection, facilitates adaptation since a new beneficial mutation sweeps to fixation in $\log(N)$ time, on average, given

that it fixes. Here one may understand the Durrett-Schweinsberg model of recurrent sweeps as approximating sweepstakes reproduction even though it does not explicitly incorporate high fecundity (understood as excess reproductive capacity in the absence of selection). Clearly the occurrence of beneficial mutations that with a probability that is independent of the population size sweep to fixation in $\log(N)$ time (the Durrett-Schweinsberg model) is what drives rapid adaptation. We have clarified the text at line 643.

Response to Reviewer # 2

1. The authors highlight, in a simulation-augmented review, recent as well as ongoing work on modelling and combining sources of sweepstake reproduction, via high fecundity-high mortality scenarios (type-III survivorship) with or without selection acting as well as as via so-called selection sweepstakes, where positive selection acts on a faster timescale than non-selective neutral evolution. I find the topics picked interesting for a broad audience, both for applied and more theoretical geneticists, and each aspect is generally described well. However, the review would read better with a framework at the end of the introduction explaining the structure and purpose of the following sections, and I object to presenting unpublished results in detail in Section 2. I will explain these points in detail below.
   Author response: We have added an outline of the paper at line 156. We have removed references to the unpublished work of Chetwyn-Diggle et al, and of Dahmer and Eldon.

2. As I see it, Section 2 introduces the models of random sweepstakes later used to assess the interplay between selection and random sweepstakes (two corrections to Schweinsberg's reproduction model whose genealogies are approximated by a Beta coalescent to make this reproduction model more biologically realistic). Then, the authors use these models to assess the impact of random sweepstakes on the allele trajectories under positive selection, including scenarios with additional bottlenecks and looking at diploid variants of said models, in Section 3. This is followed up by Section 4, which, somewhat unlinked to the sections before, presents selective sweep- stakes, the corresponding coalescent model and a review of the results of a recent preprint [Arnason et al 2022] finding strong evidence that selective sweepstakes shape the genetic diversity in Atlantic cod. The purpose of each section does only become clear after thorough reading, which in my opinion makes understanding the paper unnecessarily complex. I'd suggest that the authors add a roadmap for the following sections, so saying something like "We want to assess how positive, monogenic selection works when acting on a population undergoing random sweepstake repro- duction. In Sect. 2, we introduce the models of random sweepstakes we want to use for this. Section 3 then reviews previous studies going into that direction and assesses positive selection under various scenarios of random sweepstake reproduction. We follow this up by a further section presenting a model of selective sweepstakes, where selection alone is responsible for sweepstake reproduction". Additionally, Section 3 would

in my opinion benefit from adding further sub- sections or paragraphs with headlines that include each scenario. Finally, as indicated in my roadmap draft, the authors should clarify that their main results will look at (essentially) monogenic selection.

Author response: Thank you for suggesting a roadmap; we have added one at line 156. In it we state that selection is taken as acting at a single site. We have added subsection titles to § 3.

3. I fully agree that the adjustment to Schweinsberg's model are reasonable and I do look forward to the mathematical analysis and derivation of the coa- lescent limits announced for the two upcoming papers (Dahmer and Eldon, Chetwyn-Diggle, Eldon and Etheridge). Both the formulation of these mod- els as well as the reasons behind these adjustments are well-placed within this review and should clearly remain. However, I strongly object to explic- itly present results from these upcoming papers (the mentioned coalescent limits) without them being available as a preprint or manuscript. Unless the actual results can be scrutinized by the readers, they should not be ex- plicitly stated. I welcome mentioning the two planned papers in the text, though, but without stating the results.

Author response: We agree and have removed references to the two papers, they are now referred to as "planned future work" (e.g. line 278).

4. Line 43: The term 'rapid selection' or 'rapid adaptation' is also used in the context of selection models yielding Bolthausen-Sznitman coalescent genealogies. One or two sentences highlighting similarities and differences to your definition including 1-2 links to the literature (or appropriate review paper) would be appreciated

Author response: We believe the reviewer is referring to (Desai et al., 2013; Neher and Hallatschek, 2013) whose ideas for the effect of selection in a haploid population are made rigorous in (Schweinsberg, 2017). In this context we would also like to mention (Etheridge and Penington, 2022) who work with diploid populations where selection acts on pairs of gene copies (genotypes) rather than directly on gene copies (allelic types) as in haploid populations, and obtain the Kingman-coalescent. We briefly refer to these papers at line 573, describe the selection models as accumulations of positive mutations, and then say how our models are different in focusing on one mutation at a time.

5. Line 63: please add assumptions on LD between loci for these results
Author response: Done at line 73

6. Line 114: Rather say that a signif. proportion of all survivors comes from these winners?
Author response: Done!

7. Line 151: The double bracket "))" should be avoided
Author response: fixed! The text has been moved to Supporting Information.

8. L 151: Say "Correspondingly, when looking at genealogies backwards in time,"? *Author response: Done! at line 212*

9. L183: This is a bit technical, can you rephrase with less details in words? e.g. "Given N large enough, if each (haploid) individual has on average more than 1 offspring, the model essentially ensures that there will be enough juveniles to sample from"?
Author response: Done! at line 193

10. L 209: I'd emphasise further that this indeed is specific to Schweinsberg's model - other approaches, e.g. via modified Moran models as in [4] (used in [3]), may lead to less problematic scalings (N^α). Not saying at all that these models are better, just that this issue is somewhat model-specific.
Author response: Is indeed! Done at line 283

11. L 272++: In this and the next paragraph, in my opinion you should also discuss (Der et al., 2011).
Author response: the nice work of (Der et al., 2011) is now briefly mentioned at line 347 on line 366; note that the main text has a limit on total number of words (8000 we believe); in addition, we are asked to aim our discussion towards a broad audience, and the work of (Der et al., 2011) is quite technical. However the reviewer is right that the approach of (Der et al., 2011) might be applicable; in § S1 we briefly discuss the framework of Der et al. (2011)

12. L 289: Only for γ close to 0 the same issues than for Schweinsberg's model w. α close to 1 should appear. This is in my opinion not necessarily the case.
Author response: when identifying limiting generators one simply applies the appropriate timescaling to identify a limit, however that may require applying an unrealistic timescale, and one would want to work with a model where particular parameter values or timescalings do not cause problems.

13. L 294: Please elaborate a bit, e.g. recite/summarise one essential or representing result from (Eldon and Stephan, 2018)
Author response: we have added further clarifications at line 371

14. L 326: Note here that the simulations are based on your own code and point to your code availability section
Author response: Done at line 418

15. L 355 : Is the viability weight again e^s ? In any case, state/repeat it here
Author response: yes, and done at line 438; at line 440 we do state that how weights are assigned has already been described

16. L 446 : Make a bit of a segue here, e.g. that also selection alone may cause sweepstakes, to be modelled by specific multiple merger coalescents, which you will now describe. I would also again point to the Bolthausen-Sznitman coalescent as another model to capture a different kind of selective sweepstakes, more akin to clonal interference resp. pooling of many positively selected mutations. Quick mentioning of the model and its biological basis, telling that you don't focus on it and a pointer to the literature (or to your

introduction if you referenced there) is sufficient.
Author response: Done at line 573;

17. L 506 : typo, should be ''selective''
Author response: Nice catch! Fixed at line 643

18. L 511 : Which new models?
Author response: clarified at line 648

19. Fig 1-8: Please provide a colour legend/description
Author response: Now clarified at line 418; for each panel the trajectories shown are obtained under identical conditions; in some panels the trajectories crisscross and using the same color and line type for all of them might make it hard to distinguish between them

Response to Reviewer # 3

1. Overall, the manuscript is well written, and mostly well explained, and is of interest for this special issue of Mol Ecol. The authors present convincing theoretical evidence for the potential importance of sweepstake reproduction to fasten adaptation. My main comments before the publication of the manuscript can be recommended are as follows: 1) the readability of the manuscript can be improved to target a biological audience (at times reorder some parts of the arguments), and 2) an in depth description of some results regarding analysis of cod data should be included if these are essential in part 4. Please find below the detailed suggestions as they come.
Author response: We have revised the manuscript with biologists in mind; the cod manuscript (Árnason et al., 2022) became available on bioRxiv on June 16, 2022 (https://doi.org/10.1101/2022.05.29.493887) and we refer the interested reader to it

2. 1) In my opinion the authors overstate in the abstract the significance of their work on fitting selective sweeps to cod data because as far as I can see the manuscript Arnsasson et al. 2022 is not yet published or even available on BioRxiv? (see abstract lines 12-14 and part 4). If this paper provides key novel evidence, it should be 1) available for readers, and 2) I would suggest the authors to give more details on the results. For example, what genome size of cod, how many genomes were sequenced, which type of sweep detection was done, how many sweeps are found, what are the size of the sweep regions, how was the demographic history inferred,... Without such details we have to accept the conclusions of the authors on this important question. Such details could be given as a box focusing on cod data analysis for example. Similarly, the papers Dahmer and Eldon and Chetwyn-Diggle et al. are also not available nor published. If some results are important, these should be explicitly written and the papers available. For these three unpublished studies, the authors should tone down the claims in this review as these results cannot be assessed by the readers. If the results are unpublished, they should be remove from the reference list.

Author response: see our response to Comment #1; unfortunately the doi for the manuscript does not show up in the reference list, however bioRxiv does have a search engine, and we claim any internet search engine worth its salt would have found the cod manuscript as it became available on bioRxiv on June 16, 2022; we now provide a brief overview of the analysis and the results ; unfortunately the Dahmer & Eldon and the Chetwyn-Diggle et al papers are not yet available so we have removed them from the reference list, and refer to them as "planned future publications"

3. 2) Section 2: the argument and description are not easy to follow for a biological readership. I would suggest to move the lines 175-187 at the beginning of part 2 to give first some intuition on the reproduction mechanism, and then move to diffusion and coalescent description. Line 183: it would be nice to give a biological intuition to this property from the Schweinsberg 2003 model. Lines 206-208: Does this mean that the model is unrealistic in its fundamental description or incomplete to explain diversity in natural populations?
Author response: we have rearranged § 2 as suggested, there is also an additional clarification at line 98; the relevant property of the Schweinsberg model is now explained at line 193; in particular, we claim that not allowing for small families only (alpha > 2) as we do in Eq (4) leads to Eq (2) with the exponent 1/(alpha - 1) which makes the original Schweinsberg model incomplete in describing genetic variation

4. 3) Part 3: I may have missed the biological explanation of the definition of sudden random bottlenecks. If I understand correctly, the bottleneck lasts one generation, but can happen at any generation? How do we measure the strength and severity of bottleneck which are the important parameters defining genetic drift, if the time (length) of population reduction is a stochastic process? If I understand correctly, every simulation/trajectory shown does not exhibit the same bottleneck model, which would make difficult to compare between trajectories and likely adds some variance in the process? Would it not be easier to have a fixed time and severity of the bottleneck for all simulations?
Author response: A bottleneck can happen in any generation. At the start of a generation we toss a coin with probability b of a bottleneck, and if a bottleneck happens we sample a fixed number B of individuals to survive the bottleneck and produce juveniles, so for fixed values of b and B the fixation trajectories experience the same intensity and severity of a bottleneck effect and so are fully comparable. In our framework your idea of a fixed time for a bottleneck would correspond to changing the carrying capacity denoted by N, and we have elected to keep N fixed. In a way our model corresponds to (Eldon and Wakeley, 2006) which can be seen as a model of randomly occurring bottlenecks with immediate recovery; in our framework how quickly a population recovers from a bottlenck depends on the reproduction parameters α and the upper bound $u(N)$. We have added an exaplanation at line 448

5. A key question in part 3 is to understand what is the Ne.s selection coefficient? Can the authors give a formulae or provide some intuition (scaling difference between census size and Ne)? They give the value of s which is very high, but do not provide Ne.s though this is the relevant factor (as stated

in the introduction lines 54-55). It is thus difficult to compare these results to a Wright-Fisher model with selection as the definition of Ne under the beta-coalescent is not trivial.

Author response: We claim (to be verified in a planned future publication) that the effective size N_e (the coalescent timescale) of our mixture model in Eq (4) is of order $O(N)$ where we take N as the population size, so $N_e s$ would be $\gg 100$ or more than sufficient for detection of sweeps as stated on line 66; however, note that keeping the cutoff at N and taking $\alpha > 2$ the model behaves like a Wright-Fisher model (in allowing only small families with high probability) without changing the effective size, so $N_e s$ would then be comparable between models with and without random sweepstakes; in addition, we compute viability weights for the juveniles based on s, not $N_e s$ (see at line 403 and Eq (S11) in § S3)

6. At the end of this part, it would be nice to get ideas how these new results on dominance and/or bottlenecks are important / relevant for studying genome data in different species?
Author response: We have added a brief discussion at line 485

7. 4) Part 4: Besides my criticism in point 1 above regarding the results in Arnasson et al., I am confused by the arguments invoked here that selective sweeps explain patterns of diversity in the cod data. Do the authors mean selective sweeps in addition to a neutral sweepstake reproduction, or selective sweeps under a neutral Kingman coalescent model?
Author response: Selective sweepstakes is understood here as the mechanism invoked in (Williams, 1975) to describe the evolution of highly fecund populations; the idea is that in every generation individuals in a highly fecund population produce a huge number of juveniles and thereby introduce lots of new variation into the population through recombination and mutation, some of this new variation will be highly fit and survive to reproductive stage. In a way this is a form of fluctuating selection, where regularly new variants become the fittest types. To our knowledge there is no rigorous mathematical formulation of selective sweepstakes. However, one may view the Durrett-Schweinsberg model of recurrent sweeps of a new mutation each time arising in a population evolving according to the classical Moran model as a first approximation of selective sweepstakes, where new variants regularly sweep through the population evolving according to the Wright-Fisher (or the Moran) model. High fecundity in the sense of excess reproductive capacity in the absence of selection, as formulated for example in Eq (1) or Eq (4), is not incorporated into the Durrett-Schweinsberg model. Thus, the resulting coalescent derived from the Durrett-Schweinsberg model is the Kingman-coalescent "interrupted" by multiple mergers due to selective sweeps. Thus, what one would like to do is deriving a coalescent from a population model explicitly incorporating high fecundity and random sweepstakes as formulated in Eq (4) (or something similar) and experiencing recurrent positive mutations as in the Durrett-Schweinsberg model. Thus the resulting coalescent from such a model might be a mixture of Xi-coalescents (simultaneous multiple merger coalescents) due to random sweepstakes and due to sweeps. This is discussed in (Árnason et al., 2022) and we refer the interested reader to it.

8. Do the authors mean that selection is pervasive in the genome?
Author response: Yes, the analysis of the Atlantic cod data strongly indicates that positive selection is pervasive in the genome(Árnason et al., 2022).

9. Is it possible to disentangle between selection under sweepstake and selection under the Kingman model with complex demography?
Author response: Short answer: we don't know. Long answer: This is a key question motivating further investigation of models of random and selective sweepstakes; in (Árnason et al., 2022) there is extensive analysis of neutral Kingman-coalescent with complex demography, but we simply don't have the coalescent models to rigorously investigate this question. Deriving these models is a hard mathematical task.

10. Are these patterns robust to variation in recombination in the genome or even variation in mutation rate along the genome? If I understand correctly, would the authors expect that the genome-wide SFS in cod genomes is U-shaped because there are many (how many?) selective sweeps in the genome, and each has a limited amount of LD around it (is recombination is very high)? Or could one expect that few selective sweeps are sufficient if there is low recombination, as recombination efficiency should be decreased under the high fecundity model? Some pointer to key statistics, expectations versus observations in data would be great here.
Author response: These questions are addressed in (Árnason et al., 2022); the U-shaped SFS is consistently observed across the genome, and across coding, non-coding regions etc. The Durrett-Schweinsberg model necessarily assumes high recombination, and linkage disequilibrium statistics do indicate that there is high recombination in Atlantic cod. We argue that the pervasive U-shaped site-frequency spectra are due to many sweeps in the genome; it is difficult to give a precise estimate of the number of sweeps since the multiple-merger coalescent we fit to the data (Eq (S6)) has only one composite parameter for the strength of selection, the mutation rate, and the recombination rate(Durrett and Schweinsberg, 2005a). However, it appears that sweeps are pervasive in the genome and occur frequently relative to the coalescent timescale. On page 17 in (Árnason et al., 2022) there is discussion about the number of sweeps. There it is argued that each sweep, due to high recombination, may only affect a few percent of a chromosome, and each chromosome may be affected by a sweep every 23-50 years (recall that each sweep takes only $\log(N)$ generations on average to complete).

11. 5) The final comment regards lines 513-514 on framing analyses of sweepstake signatures in the genomic context and widen the scope of applicability beyond cod data. Maybe mentioning several species where sweepstake occur could help to point to directions of species with high/low recombination and pathogens/parasites or other organisms than cod, some examples were proposed in Tellier and Lemaire 2014, while the Menardo et al. 2020, and Jensen papers on viruses do not deal with species undergoing recombination. Along those lines, maybe the part 4 /conclusion could speculate on the distribution of selective sweeps along the genome if we are in a random sweepstake, or selective sweeps. My guess is that selective sweeps will occur preferentially in coding regions (or affecting regulatory regions), while random sweepstakes could occur anywhere in the genome especially in non-coding regions for species with large genome (and large proportions of non-coding DNA). Would this be a good expectation? Can one use the SFS of coding versus non-coding (or synonymous versus non-synonymous sites) to disentangle these hypotheses? Furthermore it seems essential to test

variation in recombination along the genome and correlation with LD of
sweeps / signatures of random sweepstake. Could the authors provide some
recommendations for testing the theory: number of samples, methods to
be used, which species and expectations are expected (to extend the early
predictions of Tellier and Lemaire 2014?). This can be also done as a
Box for example.

Author response: Random sweepstakes do affect the whole genome. Figure 5 on page 23 in (Árnason et al., 2022) shows estimates of the parameter of the Durrett-Schweinsberg coalescent for introns, exons, etc, and there is a clear difference in the estimates depending on which parts of the data are analyzed; however we recall that it is a composite parameter. As we don't have a model of recurrent selective sweeps in a background of random sweepstakes it is difficult for us to make any predictions. Providing recommendations for how to test the theory appears outside the scope of our work; there are a number of methods available for comparing complicated models (e.g. ABC-based methods or for example the diCal2 *package if one has phased data, and a recent SMC- and neural networks-based approach on bioRxiv* https: //www.biorxiv.org/content/10.1101/2022.09.28.508873v1*) The analysis in (Árnason et al., 2022) is extensive so one option might be to check the approach there. At line 103 we do list some highly fecund species that are candidates for sweepstakes reproduction.*

Sweepstakes reproduction facilitates rapid adaptation in highly fecund populations

Bjarki Eldon[1]
Wolfgang Stephan[2]

Running title: Sweepstakes reproduction facilitates adaptation

Abstract

Adaptation enables natural populations to survive in a changing environment. Understanding the mechanics of adaptation is therefore crucial for learning about the evolution and ecology of natural populations, and for better conservation and management of natural resources such as fish stocks. In this review we focus on the impact of random sweepstakes on selection in highly fecund haploid and diploid populations partitioned into two genetic types, with one type conferring selective advantage. For the diploid populations we incorporate various dominance mechanisms. Furthermore, we assume that the populations may experience recurrent bottlenecks. In random sweepstakes the distribution of individual recruitment success is highly skewed, resulting in a huge variance in the number of offspring contributed by the individuals present in any given generation. Using extensive computer simulations, we investigate the joint effects of random sweepstakes, recurrent bottlenecks, and dominance mechanisms on selection. In our framework, bottlenecks allow random sweepstakes to have an effect on the time to fixation, and in diploid populations the effect of random sweepstakes depends on the dominance mechanism. We also analyze selective sweepstakes which are well approximated by recurrent selective sweeps of strongly beneficial

[1]Institute of Evolution and Biodiversity Science, Natural History Museum Berlin, Germany
corresponding author : Bjarki.Eldon@mfn.berlin
[2]Institute of Evolution and Biodiversity Science, Natural History Museum Berlin, Germany
Department of Biology, University of Munich, Germany
December 9, 2022

allelic types arising by mutation. We demonstrate that both types of sweepstakes reproduction may facilitate rapid adaptation (as defined based on the average time to fixation of a type conferring selective advantage conditioned on fixation of the type). However, whether random sweepstakes cause rapid adaptation depends also on their interactions with environmental factors (such as bottlenecks) and genetic mechanisms (e.g. dominance mechanisms). Finally, we review a case study in which a model of recurrent selective sweeps is shown to essentially explain population genomic data of the highly fecund Atlantic cod, with implications for studying the evolution and ecology of highly fecund populations across domains of life.

1   Introduction

Many instances from the natural world show that evolutionary adaptation may occur quite rapidly. Well-known examples of rapid adaptation in response to environmental changes include color variation in guppies (Reznick, 2011), field mice (Vignieri et al., 2010) and peppered moth (Cook et al., 2012), insecticide resistance in Drosophila (Daborn et al., 2002), beak size changes in Darwin's finches(Grant and Grant, 2020), and limb development in Anolis lizards (Losos, 2009). The genetic architecture underlying these phenotypic traits ranges from a few genes of major effect as in the peppered moth (van't Hof et al., 2011) to highly polygenic systems of very small effects at individual sites such as human height (Turchin et al., 2012), although the effects of polygenic adaptation on human height appear to be overestimated (Berg et al., 2019; Sohail et al., 2019).

2

Mirroring this wide range of genetic architectures the evolutionary genetic models that have been proposed describe adaptation at a single locus (or very few loci) to polygenic adaptation involving numerous sites. Best known are the models for single loci. Clearly, very strong positive directional selection at a single locus may explain fast adaptation, such as in the case of peppered moth (Haldane, 1924). Haldane's deterministic model has been extended in several directions to make it suitable for data analysis. The extension proposed by (Maynard Smith and Haigh, 1974) for studying genetic hitchhiking (selective sweeps) is most valuable.

On the other hand, polygenic adaptation caused by a large number of weakly selected loci of small effects is not nearly as well studied as the case of strong positive selection leading to selective sweeps (Stephan, 2019). Interest of population geneticists in this type of selection was only very recently evoked by (Pritchard et al., 2010) and (Pritchard and Di Rienzo, 2010). These authors predicted that – in contrast to selective sweeps – allele frequencies may change by small amounts when a large number of genetic loci of minor effect sizes govern a phenotypic trait, but it was unclear whether such polygenic selection can explain rapid adaptation defined as follows.

How do we quantify 'rapid adaptation'? To define rapid adaptation, we assume that a population is at equilibrium when a sudden change of the fitness optimum occurs (e.g. due to a sudden shift in the environment) such that the optimum is placed to another value. If a trait is controlled by a single gene, the rate of response to selection may be quantified by the inverse of the time to fixation of a selected type given that it will go to fixation. For a diploid random-mating population of size N (assuming no dominance) evolving according to the Wright-Fisher model, the mean time to fixation (conditional on

3

fixation), taking $s > 0$ as the selective advantage of the genotype homozygous for the beneficial type relative to the homozygous wild type, is approximately $4\ln(2Ns)/s$ if $2Ns$ is sufficiently large (van Herwaarden and van der Wal, 2002). For rapid adaptation to occur $2Ns$ must be at least 100 (i.e. of about the same order of magnitude as necessary for the detection of hard selective sweeps; (Stephan, 2019)).

In the case of polygenic evolution rapid adaptation may be defined in a similar way. In this case the response of a phenotypic trait to selection may be quantified by the rate at which the population mean of the trait reaches the new fitness optimum after a sudden environmental shift. Using deterministic population genetic models of polygenic adaptation that have been analyzed in the past ten years (in particular, the deterministic model of (de Vladar and Barton, 2014), this rate can be calculated (assuming linkage equilibrium). Jain and Stephan (2015, 2017a,b) and Stephan (2016) developed formulas for the case when the effect sizes of the alleles at most loci are large relative to a scaled mutation rate such that directional selection at each of those loci is very strong, and for the opposite parameter range in which most loci involved have small effects and hence exert weak selection. While the first case is closely related to monogenic adaptation discussed above (Jain and Stephan, 2017b), in the second case the new optimum is approached exponentially (already derived by Lande (1976)), if the shift of the optimum is not too large relative to the equilibrium genetic variance. Importantly, if the effect sizes of the loci are exponentially distributed, the equilibrium variance is proportional to the number of loci (with small effects) governing the trait. Thus, if the number of loci is sufficiently large, rapid adaptation through polygenic selection may occur. Yet, in populations of finite size (in particular bottlenecks) simulations based on the classical Wright-Fisher model suggest

4

that genetic drift slows down the speed of polygenic adaptation to some extent(John and Stephan, 2020).

Here we extend a population genetic model of adaptation with Wright-Fisher drift to models of genetic reproduction that strongly deviate from the standard model of genetic drift. In particular, we consider the impact of random sweepstakes on adaptation in highly fecund populations. In random sweepstakes the distribution of individual recruitment success is assumed to be highly skewed. In contrast to the Wright-Fisher model, this results in a huge variance in individual recruitment success (the number of offspring). We investigate here in which way and to what extent this type of genetic drift affects polygenic adaptation.

Natural highly fecund populations are diverse and widely found(Eldon, 2020). By 'high fecundity' we refer to the ability of organisms to produce numbers of juveniles (potential offspring) at least on the order of the census population size. The evolution of a population over one generation is seen as each individual (or pair of individuals in diploid populations) independently producing a random number of juveniles according to a given model (i.e. probability distribution, see e.g. Eq (4) or Eq (1)); from the total pool of juveniles a given number of them is sampled (uniformly at random and without replacement) to form a new set of reproducing individuals. These include broadcast spawners such as the Antarctic limpet (*N. concinna*), Atlantic cod, Japanese sardines, crop-infesting fungi, and corals (Agrios, 2010; Árnason, 2004; Árnason and Halldórsdóttir, 2015; Barfield et al., 2022; Niwa et al., 2016; Vendrami et al., 2021). Even viruses may be classified as being highly fecund (Irwin et al., 2016; Timm and Yin, 2012). Broadcast spawners produce huge numbers of juveniles to counter the high mortality among them (i.e. Type III survivorship).

A central question regarding highly fecund populations is if the recruitment dynamics, or the distribution of individual recruitment success (the offspring number distribution) may be characterized by 'sweepstakes reproduction', or a highly skewed offspring number distribution. In this context it is important to understand how sweepstakes reproduction comes about. The Wright-Fisher model may be viewed as a model of high fecundity, where *every* individual produces a huge number of potential offspring (e.g. gametes), or at least an order of magnitude larger than the population size (see e.g. (Der et al., 2011)). Then, assuming the population size is much smaller than the number of juveniles produced by each individual, sampling juveniles uniformly and without replacement to form a new generation of individuals is well approximated by the surviving offspring sampling a parent with replacement, the sampling mechanism of the Wright-Fisher model. This mechanism does not lead to sweepstakes reproduction, since the number of surviving offspring from any given individual will be negligible relative to a large total population size. High fecundity cannot on its own produce sweepstakes reproduction. The key ingredient for sweepstakes reproduction is a mechanism that turns high fecundity into a skewed offspring number distribution, or a skewed individual recruitment success. An important open question in evolution and ecology is to infer sweepstakes reproduction in highly fecund populations, and to identify the actual mechanism of sweepstakes reproduction given evidence of it.

Since broadcast spawners characterized by Type III survivorship are highly fecund, it is plausible that the reproductive output among broadcast spawning individuals may be skewed, in the sense that occasionally, at any given time, a significant proportion of surviving offspring come from a few parents. The skew in reproductive output generated in this way without involving selection has been named 'sweepstakes reproduction', and

6

has been claimed to 'play a major role in shaping marine biodiversity'(Hedgecock, 1994; Hedgecock and Pudovkin, 2011). We will refer to this type of sweepstakes reproduction in a highly fecund population as 'random sweepstakes'(Árnason et al., 2022). Broadcast spawning and Type III survivorship combine to generate the possibility for a few lucky individuals to produce a significant number of surviving juveniles through chance matching of reproduction with favorable environmental conditions, thus being a mechanism turning high fecundity into sweepstakes reproduction.

Sweepstakes reproduction may also be generated through natural selection. In this mechanism juveniles produced at any given time are seen as having to pass through independent selective filters during development from earliest juvenile stage to reproductive age, with the result that the genetic constitution of the surviving juveniles is, on average, different from that of non-surviving juveniles(Williams, 1975). New recombined genotypes are continuously generated and carried to high frequency in a population chasing an ever-changing optimum. We will refer to this type of sweepstakes reproduction as 'selective sweepstakes', in which natural selection acts as the mechanism turning high fecundity into sweepstakes reproduction(Árnason et al., 2022).

Recruitment dynamics of natural populations, in particular the distribution of individual recruitment success, are central to the mechanisms shaping genetic diversity. Improved understanding of recruitment dynamics is therefore required for illuminating the ecology, population connectivity, local adaptation, and resilience of natural populations, for better conservation and management of fish stocks, and for much needed further development of population and evolutionary genetic theory(Botsford et al., 2001; Cowen and Sponaugle, 2009; Eldon, 2020; Fu and Li, 1999; Gagnaire et al., 2015; Grant et al., 2016; Selkoe et al.,

7

2016; Wakeley, 2004).

Here we use simulations to investigate how random sweepstakes and recurrent bottlenecks (in a highly fecund population) affects positive selection when acting at a single site. In § 2 population models of random sweepstakes are introduced. In § 3 the simulation results are presented and discussed. A case study of population genomic data of Atlantic cod(Árnason et al., 2022), which found a good fit of the data to a population model of recurrent sweeps(Durrett and Schweinsberg, 2005a), seen as approximating selective sweepstakes, is reviewed in § 4. A brief conclusion listing the main results from the simulations and important remaining follow-up projects is presented in § 5. In Supporting Information we give a brief summary of the mathematical formulation of random sweepstakes, and some further examples of the effect of random sweepstakes, recurrent bottlenecks, and dominance mechanisms on selection.

2 Modeling random sweepstakes

A natural model of random sweepstakes describes the probability distribution of the random number of juveniles, or potential offspring, contributed by each individual (see Eq (1)) in any given generation. From the pool of juveniles a given number is sampled without replacement to form a new set of reproducing individuals(Schweinsberg, 2003). The probability distribution for the random number of juveniles produced by a given individual in the Schweinsberg (2003) model is given by

$$\lim_{x \to \infty} Cx^{\alpha} \mathbb{P}\left(X_1 \geq x\right) = 1 \tag{1}$$

8

where $C > 0$ is a constant of proportionality (i.e. ensuring that the limit in Eq (1) is one) ₁₇₄

and $\alpha > 0$ is a constant determining the skewness of the distribution (Schweinsberg, ₁₇₅

2003). The formulation in Eq (1) should be understood as specifying how the probability ₁₇₆

of producing at least x juveniles behaves for very large x (at least on the order of the pop- ₁₇₇

ulation size). In a population evolving according to Eq (1) most individuals will produce ₁₇₈

few (relative to the population size) number of juveniles and so a few surviving offspring; ₁₇₉

occasionally however a single individual (the probability of two or more individuals each ₁₈₀

producing a large number of juveniles will be negligible in a large population) will produce ₁₈₁

a large (relative to the population size) number of juveniles and so a significant number ₁₈₂

of surviving offspring (illustrated in Figure 1b; in Figure 1a we illustrate for comparison ₁₈₃

the evolution of a haploid population over one generation according to the Wright-Fisher ₁₈₄

model). The model described in Eq (1) therefore corresponds well to random sweepstakes ₁₈₅

occurring in a broadcast spawner evolving according to Type III survivorship; individuals ₁₈₆

must have the capacity to produce huge numbers of juveniles to counter the high mortality ₁₈₇

among the juveniles, and once in a while a lucky individual matches reproduction with ₁₈₈

favorable environmental conditions so a significant fraction of the lucky individual's juve- ₁₈₉

niles survive. How often such large families occur, and how large they will be, depends ₁₉₀

on the value of α in Eq (1) or (3) determining the skewness of the probability distribution ₁₉₁

for the number of juveniles produced by any given individual; the smaller α is the higher ₁₉₂

the chance of producing many juveniles. From the total pool of juveniles, assuming there ₁₉₃

are enough of them (at least N of them in a haploid population of size N), which is almost ₁₉₄

guaranteed in a large population provided each individual produces more than one juve- ₁₉₅

nile on average(Schweinsberg, 2003), we then sample N juveniles uniformly at random ₁₉₆

9

and without replacement to form a new set of reproducing individuals. The parameter α [197] in Eq (1) is the quantity determining how quickly the probability of producing at least x [198] juveniles decays as x increases. In the case $0 < \alpha < 2$ large families occur often enough and [199] are large enough to affect the evolution of the population. That means that the ancestral [200] process (the process tracking the random ancestral relations of sampled gene copies) is in [201] the domain of attraction of a particular example of a multiple-merger coalescent (which, [202] in contrast to the Kingman-coalescent, admits mergers of at least three ancestral lineages) [203] generally referred to as the Beta$(2 - \alpha, \alpha)$-coalescent(Schweinsberg, 2003). [204]

Population genetic models incorporating random sweepstakes are in fundamental ways [205] different from the classical Wright-Fisher model. For example, the frequency process of [206] a genetic type segregating in a population evolving according to random sweepstakes is [207] in the domain of attraction of a jump diffusion where the process, in addition to evolving [208] according to the well known Wright-Fisher diffusion, admits discontinuous jumps(Birkner [209] and Blath, 2009). The jumps correspond to the occurrence of large families involving a [210] number of copies of the type being tracked (a brief overview of the mathematical formula- [211] tion is given in § S1). Correspondingly, when looking at genealogies of samples, models of [212] random sweepstakes are in the domain of attraction of multiple-merger coalescent models [213] (a coalescent is a probabilistic description of the random ancestral relations of a sample [214] of gene copies from an arbitrarily large hypothetical population evolving according to a [215] given model)(Berestycki, 2009; Eldon and Wakeley, 2006; Huillet and Möhle, 2011, 2013; [216] Möhle, 2011; Pitman, 1999; Sagitov, 1999; Sargsyan and Wakeley, 2008; Schweinsberg, [217] 2003), where a random number of ancestral lineages of a sample of gene copies merges [218] at any given time(Donnelly and Kurtz, 1999; Pitman, 1999; Sagitov, 1999; Schweinsberg, [219]

2000). Coalescent theory provides a framework for the development of powerful inference methods(Wakeley, 2007). Multiple-merger coalescent processes predict patterns of population genetic data that are different from predictions of the classical Wright-Fisher model and similar models(Birkner et al., 2013b; Blath et al., 2016). To identify random sweepstakes in natural populations one can therefore apply coalescent-based inference to population genetic data(Birkner and Blath, 2008; Birkner et al., 2011, 2013c; Eldon, 2011, 2016; Eldon et al., 2015; Freund and Siri-Jégousse, 2021; Koskela, 2018; Koskela and Berenguer, 2019). In contrast to multiple merger coalescents, the gene genealogy of a sample from a population evolving according to the Wright-Fisher model, or similar model (where 'similar' refers to certain conditions on the offspring number distribution(Möhle and Sagitov, 2001; Sagitov, 1999)) is described by the Kingman-coalescent, in which no more than two ancestral lineages merge each time(Berestycki, 2009; Kingman, 1982a,b,c; Tajima, 1983). Intuitively one would expect a multiple-merger coalescent to describe gene genealogies under random sweepstakes, since whenever a large family occurs, it will involve a number of the ancestral lineages with non-negligible probability in a large population. Correspondingly, looking forward in time, a number of copies of the genetic type being tracked will be involved in the large family event, leading to a jump in the type frequency process. Figure 1 records a simple illustration of the difference between the Wright-Fisher model and random sweepstakes for haploid populations. The occational occurrence of a large family as shown in Figure 1(b) can induce jumps in the frequency of the type being tracked. Our aim is to investigate, given that a population evolves according to random sweepstakes, how does that affect the fate of advantageous mutations, and by implication adaptation (Figure 2).

Even though the model in Eq (1) seems natural, it has two main drawbacks. One is that individuals are assumed to be able to produce arbitrarily many juveniles. Although some organisms are extremely fecund, the assumption of unbounded fecundity is unrealistic. The second assumption involves the scaling of time used to pass to continuous-time limit process as population size tends to infinity In passing to a coalescent limit (in probability theory this means convergence in a certain sense, in our case in terms of finite-dimensional distributions), one scales time with a quantity usually denoted c_N, which is the probability that two distinct gene copies sampled at the same time derive from the same parental copy in the previous generation. For a haploid population of size N evolving according to the Wright-Fisher model, $c_N = 1/N$ (so that N generations correspond to one coalescent time unit). In the random sweepstakes model given in Eq (1) with $1 < \alpha < 2$, we have $c_N \propto N^{1-\alpha}$. This means that, for estimates of α close to one (e.g. (Árnason and Halldórsdóttir, 2015)), an unrealistically high population size is required to recover the observed genetic variation(Eldon, 2020). To see this, we consider the expected number of segregating sites (see Eq (S12) in § S5) of a sample of size n of a non-recombining contiguous chromosome segment of length L, $\mathbb{E}[B(n)]$ denotes the expected tree size (i.e. the expected size of the random gene genealogy connecting the n sampled gene copies under a given population model) with time measured in coalescent time units, and μ the per site per generation mutation rate. We can safely take $\mathbb{E}[B(n)] = O(1)$, and then we obtain (see § S5), assuming we observe m segregating sites,

$$
N \approx \left(\frac{1}{\mu} \frac{m}{L} \right)^{\frac{1}{\alpha-1}} .
\tag{2}
$$

Assuming around one percent of sites are segregating, i.e. $m/L \approx 0.01$, and μ of order

12

$O(10^{-8})$ as for Pacific cod(Canino et al., 2010), we see that for estimates of α close to one it would be difficult to recover the observed amount of genetic variation without requiring an unrealistically large population size.

To address the assumption of the Schweinsberg (2003) model regarding unbounded fecundity, one can apply an upper bound on the number of juveniles produced by an individual(Eldon and Stephan, 2018). Suppose X_1 denotes the random number of juveniles produced by a given individual, consider the mass function (i.e. the probability of producing x juveniles)

$$\mathbb{P}\left(X_1 = x\right) = \mathbb{1}_{\{1 \leq x \leq u(N)\}} \left(\frac{1}{x^{\alpha}} - \frac{1}{(1+x)^{\alpha}}\right) \frac{(1+u(N))^{\alpha}}{(1+u(N))^{\alpha} - 1}, \tag{3}$$

where $u(N)$ is an increasing positive function of N and representing an upper bound on the number of juveniles a given individual can produce. The parameter $\alpha > 0$ determines how the probability of producing x juveniles decreases with increasing x. The smaller α is the higher the probability of producing many juveniles, and $\mathbb{1}_{\{A\}} := 1$ if A holds, and zero otherwise. The model in Eq (3) is a variant of the model in Eq (1). The behaviour of the model with respect to varying α and $u(N)$ is being rigorously investigated for a planned future publication. The model in Eq (3), or some variant of it, in addition to modeling random sweepstakes (as α decreases the probability of producing many juveniles increases) can also serve as a natural, realistic, and a mathematically tractable alternative to the Wright-Fisher model (as α increases the probability of producing many juveniles decreases).

In order to address the second drawback of the Schweinsberg (2003) model regarding the scaling of time we consider a simple variant of the model in Eq (3). A similar

13

approach is also adopted in (Eldon and Wakeley, 2006; Huillet and Möhle, 2013), where $\,$285

convergence to coalescents are obtained with less problematic scalings of time. Suppose $\,$286

with probability ε_N all current N individuals produce juveniles according to Eq (3) with $\,$287

with α 'small', or an increased probability of producing many juveniles, with probability $\,$288

$1 - \varepsilon_N$ we take α 'large' (representing a decreased probability of producing many juve- $\,$289

niles); in both cases we suppose the cutoff $u(N)$ is proportional to the population size as $\,$290

population size N increases, e.g. taking $u(N) = KN$ for some finite constant $K > 0$. The $\,$291

advantages of this approach are twofold. One is that in this model ordinary reproduction $\,$292

(in which each individual produces a small number, relative to the population size, of ju- $\,$293

veniles with high probability) occurs most of the time. Occasionally (i.e. with probability $\,$294

ε_N), however, reproduction matches favorable environmental conditions, each individual $\,$295

produces juveniles according to Eq (3) with smaller α, or with an increased probability of $\,$296

producing many juveniles. In this way random sweepstakes can be modelled to be strong $\,$297

enough to impact the evolution of the population without being the overwhelming force. $\,$298

We also claim that the second advantage of the mixture model regards the scaling of time $\,$299

(recall Eq (2)) required to pass to a coalescent limit (this is also being investigated in a $\,$300

planned future work). Now we describe the mixture model we use for the simulations. $\,$301

Let $L(\alpha, u(N))$ denote the law (probability distribution) of the number of juveniles with $\,$302

mass function as in Eq (3), and take $\alpha_1 \in (0, 2)$ and $\alpha_2 \geq 2$ as fixed. The quantity α_1 $\,$303

represents an increased probability of producing many juveniles, and α_2 a decreased prob- $\,$304

ability of doing so. Assuming the cutoff $u(N)$ is at most of order N, i.e. we assume the $\,$305

fecundity is high enough to impact the evolution of the population, but not necessarily an $\,$306

unbounded fecundity, we will write the mixture-model as , where $X_1, \ldots, X_N$ denote the $\,$307

random number of juveniles produced by the current N individuals,

$$
X_1, \ldots, X_N \sim
\begin{cases}
L(\alpha_2, u(N)) & \text{with probability } 1 - \varepsilon_N \\
\\
L(\alpha_1, u(N)) & \text{with probability } \varepsilon_N
\end{cases}
\tag{4}
$$

i.e. the $X_1, \ldots, X_N$ are independent and identically distributed as specified in Eq (4). The meaning of Eq (4) is as just described, that with probability $1 - \varepsilon_N$ the number of juveniles is distributed according to Eq (3) with $\alpha = \alpha_2$ (a decreased probability of producing many juveniles) , and with probability ε_N we take $\alpha = \alpha_1$ (i.e. reproduction matches favorable environmental conditions and every individual produces juveniles with an increased probability of producing many juveniles). Similarly one can keep α fixed between one and two and randomize the cutoff $u(N)$; however we will restrict ourselves to the model in Eq (4). The model in Eq (4) is a natural way of formulating random sweepstakes in a broadcast spawner; most of the time individuals produce a small (relative to the population size) number of juveniles ($\alpha = \alpha_2 \geq 2$); occasionally (with probability ε_N) favorable environmental conditions match reproduction so that there is a higher chance of producing a larger number of juveniles ($\alpha = \alpha_1 \in (0, 2)$). Models corresponding to Eq (4) for diploid (or polyploid) populations would necessarily involve simultaneous multiple-merger coalescent processes where at least two distinct groups of ancestral lineages could merge at a time(Birkner et al., 2013a, 2018; Blath et al., 2016; Koskela and Berenguer, 2019; Möhle and Sagitov, 2003; Sagitov, 2003; Schweinsberg, 2000). The Beta$(2 - \alpha, \alpha)$-coalescent based on the original population model of random sweepstakes (see Eq (1); Schweinsberg (2003)) for both haploid and diploid (Birkner et al., 2018) populations has been implemented in the state-of-the-art simulation package `msprime`(Baumdicker et al.,

15

3   The impact of random sweepstakes on selection

The impact of selection on the evolution of a population can be measured in at least

two ways. One is with the probability of fixation, denoted $p_N(y)$ (see Eq (S10) in § S2 in

Supporting Information), of the beneficial type (the type conferring a selective advantage)

from a given number y of copies of the type, and another is with the time to fixation

conditional on fixation of the beneficial type denoted $\tau_N(y)$ (see § S2). The quantity

$p_N(1)$ is the probability of fixation of the beneficial type when starting from one copy of

the type, and $\tau_N(1)$ is the expected time to fixation of the beneficial type conditional on

fixation the type and starting with one copy of the beneficial type. We are interested in

investigating the effect of random sweepstakes on $p_N(1)$ and $\tau_N(1)$ under a simple model

of viability selection. The fixation probability can inform about adaptation, including

the occurrence of resistance to antibiotics, and about loss of genetic variation(Patwa and

Wahl, 2008). The expected time $\tau_N(1)$ is well defined since we consider a finite population,

so that the boundaries $(0, N)$ will be reached in finite time almost surely. The expected

time can inform about the rate of adaptation given that it will occur.

   As we have discussed (§ 2), models of random sweepstakes admit jumps in the genetic

type frequency. This means that classical diffusion techniques (Feller, 1951; Kimura, 1957)

are not applicable. Nevertheless, some mathematical results have been obtained on the

impact of random sweepstakes on selection(Der et al., 2012, 2011; Foucart, 2013).   One

way to approach this problem might be to identify the limiting generators (as population

size tends to infinity) of the forward-in-time process and then work with the generators as

done by Der et al. (2011) who introduce generalized Wright-Fisher models (see § S1). Consider the following simple model of random sweepstakes. Suppose a population evolves according to a discrete-time Moran model (i.e. one randomly picked individual produces offspring, and an equivalent amount of individuals perish to keep the population size constant). The distribution of the random number of offspring (V) produced by the parent at any given time is

$$\mathbb{P}\left(V = v\right) = \mathbb{1}_{\{v=1\}}(1 - \varepsilon_N) + \mathbb{1}_{\{v=\lfloor \psi N \rfloor\}}\varepsilon_N. \tag{5}$$

In Eq (5) (as in Eq (4)) ε_N can be understood as the probability of matching reproduction with favorable environmental conditions, and when that happens the parent produces $\lfloor \psi N \rfloor$ surviving offspring. Each time as many individuals as produced perish to keep the population size N constant, and $0 < \psi, \varepsilon_N < 1$(Eldon and Wakeley, 2006). Viewing time in units of $O(N^\gamma)$ generations for some $\gamma \in (0, 2]$, it can be shown that the model is in the domain of attraction of a Λ-coalescent , indeed for $\gamma = 2$ the limiting coalescent is a mixture of the Kingman-coalescent and a multiple-merger coalescent (Eldon and Wakeley, 2006). The model in Eq (5) is unrealistic in assuming that exactly the same fraction (ψ) of the population is replaced in each sweepstakes event, but it is among the simplest models of random sweepstakes, and its simplicity does facilitate some mathematical results to be obtained(Der et al., 2012; Eldon and Freund, 2018; Matuszewski et al., 2018). In particular, fixation probabilities under selection are studied for generalized Wright-Fisher processes(Der et al., 2011). Assuming random sweepstakes according to Eq (5) in a framework involving selection, conditions on the strength of selection can be identified under which fixation of the fitter of two genetic types is assured(Der et al., 2012). The results

17

on assured fixation are limit results as population size $N \to \infty$(Der et al., 2012). The 371

results of a simulation study of the effect of random sweepstakes on selection in a finite 372

haploid population, based on the model in Eq (3) and not on the mixture model in Eq (4), 373

indicate that fixation is anything but given in a finite population evolving according to 374

random sweepstakes , i.e. the fixation probability is clearly decreased as α tends to one, 375

the distribution of the number of juveniles is highly skewed, and individuals are allowed 376

to produce numbers of juveniles an order of magnitude larger than the population size 377

(Eldon and Stephan, 2018). 378

3.1   Haploid populations 379

Assuming a haploid population of constant size N evolving by random sweepstakes ac- 380

cording to Eq (3) under a simple model of viability selection, simulation results indicate 381

that both $p_N(1)$ and $\tau_N(1)$ are much smaller under random sweepstakes than under ordi- 382

nary (i.e. individuals produce small number, relative to the population size, of juveniles 383

with high probability) reproduction(Eldon and Stephan, 2018). That is, the chance of fix- 384

ation is significantly smaller under random sweepstakes, and the expected time to fixation 385

when fixation happens is shorter under random sweepstakes. One must view these results 386

in the context of the expected number of beneficial mutations in the whole population 387

over a period of time. The mixture model in Eq (4) may moderate the reduction in the 388

probability of fixation due to random sweepstakes relative to Eq (3), the model in (Eldon 389

and Stephan, 2018). Thus, if the number of beneficial variants that occur in the population 390

before the most recent common ancestor of the whole population is reached is sufficiently 391

large that a somewhat smaller chance of fixation of any one of them due to random sweep- 392

18

stakes will not significantly alter their overall effect, then one would expect that random

sweepstakes would affect the evolution of the population when new beneficial mutations

arise. Figure 2(b) records a simple illustration of the effect of random sweepstakes on

selection through the fixation of an advantageous type (for comparison Figure 2(a) illus-

trates fixation in a haploid population evolving according to the Wright-Fisher model).

To investigate the impact of random sweepstakes acting through Eq (4) on selection

we model selection as follows. Suppose a haploid population of constant size N evolves

in discrete generations. In every generation all N current individuals independently con-

tribute juveniles according to Eq (4). The population is partitioned into two genetic types,

one conferring viability weight one, and the other viability weight e^{-s}. Throughout we

start with the fitter type in one copy. Juveniles inherit the type of the parent, we exclude

mutation. Given a pool of S_N juveniles we sample independent exponentials each with the

rate of the viability weight of a given juvenile. That is, if there are S_y juveniles with viability

weight one, and S_{N-y} juveniles ($S_y + S_{N-y} = S_N$) with viability weight e^{-s}, we sample S_y in-

dependent exponentials with rate one each, and S_{N-y} independent exponentials with rate

e^{-s}. The N juveniles with the smallest exponentials survive to form the next generation

of reproducing individuals. This is a way to let selection influence the viability of each

juvenile, and has been applied in a previous investigation on the effects of random sweep-

stakes on selection(Eldon and Stephan, 2018). In the present work we model random

sweepstakes based on Eq (4), whereas earlier (Eldon and Stephan, 2018) we kept both

α and the cutoff $u(N)$ fixed. In Figures 3– 4 (see also Figures S1– S5 in Supplementary

Information) we show examples of excursions to fixation for several scenarios. We do not

aim for precise estimates of $\tau_N(1)$ or $p_N(1)$ (Eq (S10)), but rather to see main trends in how random sweepstakes modeled through Eq (4) affect $\tau_N(1)$ and $p_N(1)$. In the Figures the scale of the abscissa (horizontal axis) may vary between subplots. In each panel the colours and line types are only meant to distinguish between the trajectories; for any given panel the trajectories shown were all obtained under identical conditions. The C++ code written for the simulations is freely available on github (see § 6.1).

The size of natural populations changes in time, in particular there may be randomly occurring bottlenecks (sharp reduction in population size). For example, bottlenecks have been suggested to be an important factor in the evolution of resistance of pathogenic bacteria to antibiotics (Mahrt et al., 2021). We model bottlenecks as follows. Suppose there is a fixed upper bound on the population size, i.e. the total number of individuals in the population cannot be more than some fixed number N. We can think of this number as the carrying capacity of the environment. In any given generation a bottleneck occurs with a fixed probability. Should a bottleneck occur we sample a fixed number (denoted N_b) of individuals uniformly at random that will survive the bottleneck (the remaining individuals, the ones not surviving the bottleneck, are discarded). We then check if any of the surviving individuals contain the beneficial type; if not we stop since the beneficial type is then lost from the population. On the other hand, if the beneficial type is fixed among the surviving individuals we stop and record a fixation of the type. If the beneficial type is present among the surviving individuals but has not fixed, the surviving individuals produce juveniles according to the given model (Eq (4)). If the total number of juveniles is less than N all the juveniles are assumed to survive, in this way we allow the population to recover from the bottleneck. Otherwise we assign a viability weight (weight one to the fit

20

type and weight e^{-s} to the wild type for haploid populations; for diploid populations see 439

Eq (S11)) to each juvenile and sample N of them as described above. Write B for the event 440

a bottleneck occurs in a given generation, B^c the event a bottleneck does not occur in a 441

given generation , and define $\mathbb{1}_{\{B\}} := 1$ if B occurs, and zero otherwise. Denoting N_t the 442

population size at time t, we take $N_{t+1} = \min(S_M, N)$, where S_M denotes the total number of 443

juveniles produced by $M = \mathbb{1}_{\{B\}}N_b + \mathbb{1}_{\{B^c\}}N_t$ individuals. Thus, viability selection comes 444

into effect only when the total number of juveniles exceeds N. This way of modeling 445

bottlenecks is somewhat similar to the model in Eq (5) (see (Eldon and Wakeley, 2006)). 446

The model in Eq (5) can be seen as a model of an instantaneous bottleneck followed by 447

immediate recovery of the population. A bottleneck can happen in any generation. At the 448

start of a generation we toss a coin with probability b of a bottleneck, and if a bottleneck 449

happens we sample a fixed number B of individuals to survive the bottleneck and produce 450

juveniles. Thus, our model corresponds to the model in Eq (5) (Eldon and Wakeley, 451

2006) which can be seen as a model of randomly occurring bottlenecks with immediate 452

recovery; in our framework how quickly a population recovers from a bottlenck depends 453

on the reproduction parameters α and the upper bound $u(N)$ on the number of juveniles 454

an individual (or a parent pair in diploid populations) can produce (see Eq 3). 455

In Figure 3 we investigate the effects of randomly occurring bottlenecks on the evo- 456

lution of a haploid population evolving according to Eq (4). To recall the mechanism 457

described in Eq (4), individuals produce a small number (relative to the population size) 458

of juveniles most of the time (with probability $1 - \varepsilon_N$) but with a chance (represented by 459

ε_N), when reproduction matches favorable conditions, of producing an increased number 460

of juveniles, with carrying capacity $N = 10^6$, $\alpha_1 = 0.75$ (representing an increased proba- 461

bility of producing many juveniles according to Eq (3)), $\alpha_2 = 3$ (representing a decreased probability of producing many juveniles), cutoff $u(N) = N$ (each individual can produce at most $u(N)$ juveniles), and selection strength $s = 0.5$ throughout. The parameters that vary as shown in Figure 3 are ε_N (the probability of matching reproduction with favorable environmental conditions, so that individuals have a higher chance of producing many juveniles, i.e. $\alpha = \alpha_1$ in Eq (3)), bottleneck size B (the number of individuals surviving a bottleneck), and b the probability of a bottleneck in a given generation. Bottlenecks clearly affect the probability of fixation; the top row shows the excursions to fixation for 10^2 experiments, but the remaining panels all show excursions for 10^3 experiments except for panel (j), where there are over 16000 experiments. We emphasize that we are interested in uncovering broad trends in how random sweepstakes and bottlenecks affect selection, and we are not aiming for precise estimates of $\tau_N(1)$ and $p_N(1)$ (Eq (S10)). Figure 3 shows that if bottlenecks on average occur frequently (high probability of a bottleneck) the probability of fixation of the advantageous type ($p_N(1)$, Eq (S10)) is reduced. We claim a lower $p_N(1)$ would be expected from a high frequency of bottlenecks, since then a bottleneck will occur with high probability while the beneficial type is still in low frequency (we always start with the beneficial type in one copy), and so there is a good chance of losing the type through a bottleneck. Furthermore, if bottlenecks occur frequently but are not too severe the time to fixation is increased if the random sweepstakes are not too severe (top two rows); if random sweepstakes occur with high probability ($\varepsilon_N = 0.1$, bottom two rows) they clearly cancel out the effect of bottlenecks on the time to fixation. We estimate that strong random sweepstakes as given by the scenario in Figure 3(g-l) reduce the probability of fixation roughly tenfold compared to the case of weak sweepstakes Figure 3(a-f).

22

The fixation trajectory (or a part of it) of a beneficial type in a haploid population has been described with a logistic differential equation (LDE; (Kaplan et al., 1989; Stephan et al., 1992); see Eq (S13) in § S6). See (Schweinsberg and Durrett, 2005) for another approach to describe fixation trajectories. Figure S1 in § S2 gives examples of trajectories well approximated by a LDE even in the presence of (moderated with $u(N) = N$) random sweepstakes (Figure S1b). Frequent recurrent bottlenecks (Figure 3d–f, j–l) clearly cause significant deviations from the logistic curve; even less frequently occurring bottlenecks (Figure 3a–c,g–i) generate notable deviations from the LDE. The trajectory a given mutation travels towards fixation may inform about the shape of the site-frequency spectrum of a sample. There is pervasive U-shape of the site-frequency spectrum in genomic data from Atlantic cod(Árnason et al., 2022), and the mutations at observed segregating sites may represent mutations traveling along a fixation trajectory on their way to fixation. The trajectories of a completely dominant fit type (the heterozygote is as fit as the homozygote for the fit type, see Eq (S11) in § S3) are characterized by an extended time with the fit type at high frequency (Fig 4). This type of a trajectory indicates that mutations that we pick up in a sample showing a U-shaped site-frequency spectrum may either be mutations under positive selection traveling on this kind of a trajectory, or other mutations hitchhiking with a positive mutation; this may explain the excess (relative to predictions of the Kingman-coalescent) of mutations in high frequency in the right tail of a U-shaped site-frequency spectrum. Forward-in-time simulations (using SLiM(Haller and Messer, 2016)) of strongly beneficial dominant or semi-dominant positive mutations arising in a population evolving according to the Wright-Fisher model yielded site-frequency spectrum matching quite well the observed spectra from Atlantic cod(Árnason et al., 2022). Taken together, we believe

23

that these results show that knowledge of the shape of fixation trajectories may inform 508

about the footprint of selection in data. 509

To summarize the simulation results for haploid populations and focusing on the time 510 to fixation ($\tau_N(1)$ in Eq S9) frequently occurring bottlenecks increase $\tau_N(1)$ relative to in- 511 frequently occurring bottlenecks when random sweepstakes occur infrequently (Figure 3a- 512 f). Increasing the frequency of random sweepstakes (Figure 3g-l) counteracts the effect 513 of bottlenecks on $\tau_N(1)$. In the absence of bottlenecks random sweepstakes as modelled 514 in Eq (4) with a bound on the number of juveniles have little effect on $\tau_N(1)$ (Figure S1). 515 In the complete absence of random sweepstakes (Figure S2), increasing the frequency of 516 bottlenecks (Figure S2d-f) increases $\tau_N(1)$, however this depends on the severity of the 517 bottleneck. Bottlenecks thus allow random sweepstakes to have an effect on $\tau_N(1)$. The 518 upper bound ($u(N)$ in Eq (3)) for the number of juveniles produced by any individual 519 remains fixed at the carrying capacity. This means that when a bottleneck occurs the pop- 520 ulation size becomes smaller than the cutoff, thus increasing the chance for individuals 521 to produce a large number of juveniles relative to the population size. This is consistent 522 with previous simulation results for haploid populations of fixed size, where it was seen 523 that taking the cutoff larger than the population size shortened $\tau_N(1)$ (Eldon and Stephan, 524 2018). 525

3.2 Diploid populations 526

In addition to haploid populations we consider the effect of random sweepstakes and ran- 527 domly occurring bottlenecks on selection in diploid populations. To this end we consider 528 a diploid population of maximum size $2N$ diploid individuals (the carrying capacity). In 529

24

any given generation the current diploid individuals arbitrarily form pairs, and the pairs then independently produce juveniles according to Eq (4). The juveniles are assigned gene copies following Mendel's laws, i.e. each diploid juvenile receives one gene copy from each diploid parent. The genotype of each given juvenile then determines the viability weight as described in Eq S11 in § S3. We then proceed as previously described for haploid populations. Diploidy gives us an opportunity to investigate the joint effects of dominance mechanisms, random sweepstakes, and randomly occurring bottlenecks on $p_N(1)$ and $\tau_N(1)$ (Eq (S10)). We will consider complete dominance and incomplete dominance of the fit type as well as the case with the fit type being recessive (see Eq (S12) in § S3). In all cases the optimal genotype is the homozygous 1/1 type. In the complete dominance case it makes sense to consider 1/1 as the optimal type, since heterozygotes contain the 0 type, so while there are heterozygotes in the population there is always a chance of a 0/0 type.

In Fig 4 we compare the effects of random sweepstakes and randomly occurring bottlenecks on $p_N(1)$ and $\tau_N(1)$ defined in Eq (S10) in the case of complete dominance of the beneficial type (Eq (S12)). Random sweepstakes ($\varepsilon_N = 0.1$, bottom two rows) reduce $p_N(1)$ (Eq (S10)) only slightly. The effect of random sweepstakes on $\tau_N(1)$ is particularly noticeable in the case of a 'weak' bottleneck (bottleneck size 10^4, right column). The case of the beneficial type showing incomplete dominance as defined in Eq (S12) is investigated in §S3.2, see Figure S4.

In the case of the beneficial type being recessive (Figure S5 in § S3.3) without random sweepstakes ($\varepsilon_N = 0$ in Eq (4)), the time to fixation is longer if the bottlenecks are weak (more individuals surviving a bottleneck) (Figure S5b,d). The main effect of

25

random sweepstakes is to shorten $\tau_N(1)$, while bottlenecks tend to increase $\tau_N(1)$. Furthermore, there is a clear qualitative difference in the excursions to fixation depending on the dominance mechanism (Figures S3– S5).

To summarize the simulation results for diploid populations and focusing on $\tau_N(1)$, the effect of random sweepstakes depends on the dominance mechanism. In the absence of bottlenecks (Figure S3) random sweepstakes have negligible effect on $\tau_N(1)$ when the fit type is incompletely dominant (Figure S3a,b), but clearly reduce $\tau_N(1)$ when the fit type is either completely dominant (Figure S3c,d), or recessive (Figure S3e,f). When the fit type is dominant (Figure 4), increasing the frequency of bottlenecks increases $\tau_N(1)$ in the absence of random sweepstakes (Figure 4a-d); introducing random sweepstakes (Figure 4e-h) largely negates the effect of bottlenecks on $\tau_N(1)$. Introducing frequently occuring bottlenecks to a population with an incompletely dominant fit type (Figure S4) increases $\tau_N(1)$ in the absence of random sweepstakes (Figure S4c-d) relative to less frequently occuring bottlenecks (Figure S4a-b); and again introducing random sweepstakes (Figure S4e-h) largely nullifies the effect of bottlenecks on $\tau_N(1)$. Similarly one can compare the effects of random sweepstakes and bottlenecks on $\tau_N(1)$ when the fit type is recessive (Figure S5). As in the the case of haploid populations, bottlenecks allow random sweepstakes to have an effect on $\tau_N(1)$ since the cutoff ($u(N)$ in Eq (3)) remains fixed at the carrying capacity, so is larger than the population size when a bottleneck occurs.

4 Selective sweepstakes

We have discussed the effect of random sweepstakes on selection. Now we turn our focus on selection in the absence of random sweepstakes, and discuss possible footprint of se-

26

lection in data. For example, models of selection in haploid populations, where positive mutations have accumulated, and incorporating a form of clonal, or Hill-Robertson inter- ference(Hill and Robertson, 1966), can lead to multiple-merger gene genealogies(Desai et al., 2013; Neher and Hallatschek, 2013; Schweinsberg, 2017). In contrast, underdomi- nance of the fit type in a diploid population with two types, where the heterozygote is less fit than both homozygotes, results in the Kingman-coalescent describing the gene geneal- ogy of a sample of the fit type(Etheridge and Penington, 2022). Here, we do not focus on these models. We will be concerned with a model of selection that focuses on one selec- tively advantageous mutation at a time corresponding to our simulations of the evolution of a selectively advantageous type always starting in one copy.

Selective sweepstakes are a form of sweepstakes reproduction in which, in contrast to random sweepstakes, natural selection plays a key role. In selective sweepstakes juveniles are seen as having to pass through independent selective filters as they go through the different developmental stages on their way to reproductive age(Williams, 1975). We are not aware of a mathematical model precisely for selective sweepstakes. One may, however, view selective sweepstakes as being well approximated by models of recurrent selective sweeps(Coop and Ralph, 2012; Durrett and Schweinsberg, 2004, 2005b).

The Durrett-Schweinsberg model (Durrett and Schweinsberg, 2004, 2005b) of recurrent selective sweeps (see Fig S6), each time from a new and strongly beneficial mutation, has been shown to explain population genomic data of Atlantic cod(Árnason et al., 2022). Furthermore, current models of random sweepstakes based on Eq (1) (Birkner et al., 2018; Schweinsberg, 2003), in addition to the Kingman coalescent incorporating complex demography and background selection, do not explain the Atlantic cod data(Árnason

et al., 2022). To better understand the importance of this result we now briefly describe the Durrett-Schweinsberg model(Durrett and Schweinsberg, 2004, 2005b). Consider a haploid population of constant size $2N$ evolving according to the continuous-time Moran model, i.e. at exponentially distributed times a single individual contributes a single offspring, and another individual is removed to keep the population size constant. A new beneficial mutation occurs on a chromosome with rate proportional to $1/N$, where with probability $S > 0$ (independent of N) the new mutation will sweep to fixation in $\log(N)$ time units on average. Viewing time in N time units the gene genealogy of a neutral site on a chromosome converges (as $N \to \infty$) to a coalescent, which is a mixture of the Kingman coalescent and a multiple-merger coalescent (see Eq (S6)).

However, even though the Durrett-Schweinsberg model (see § S4) , essentially explains the U-shape (Fig 5) of the site-frequency spectrum of Atlantic cod, there are certain limitations to the model. It is essentially a haploid model, where selection acts directly on individual chromosomes rather than pairs of chromosomes in diploid individuals. Furthermore, in order to obtain a non-trivial coalescent (see Eq (S6) in § S1) , the advantage of the beneficial mutation must be of order $O(1)$ to lead to $\log(N)$ as the order of the average time it takes to sweep to fixation. Viewing time on the scale of N time units a sweep then generates instantaneous (multiple) mergers in the genealogy. However, to avoid an instantaneous merger of all the lineages, some lineages must be allowed to escape a sweep through recombination. This is a key element of the model. The short duration (on average) relative to the coalescent timescale of a sweep means that one must assume very high recombination rates in order to allow an escape during a sweep. Despite these limitations, the exceptionally good fit of the Durrett-Schweinsberg model to the Atlantic cod

28

data shows that models of recurrent and pervasive selective sweeps, and multiple-merger coalescent models, are relevant for explaining genetic diversity in highly fecund natural populations. Indeed, related models of incomplete sweeps also lead to multiple-merger coalescents, with similar predictions of genetic diversity(Coop and Ralph, 2012).

5   Conclusion

We have discussed adaptation in natural populations with a focus on highly fecund populations evolving according to sweepstakes reproduction. We have discussed two quite different mechanisms that turn high fecundity into skewed individual recruitment success. We have suggested new models of random sweepstakes that address the assumption of unbounded fecundity and the timescale issue and the related problem of recovering observed amount of genetic variation discussed above. Finally, we have used simulations to identify the main trends in how random sweepstakes affect fixation of a beneficial type under a simple model of viability selection with randomly occurring bottlenecks. The main impact of random sweepstakes is to reduce the time to fixation conditional on fixation (compared to ordinary reproduction). This suggests that random sweepstakes facilitate rapid adaptation. Yet, as in the case of polygenic selection mentioned in Introduction, our one-locus model predicts that bottlenecks may increase the time to fixation (conditional on fixation) and thus limit the speed of adaptation.

We give examples of fixation trajectories to understand how a mutation sweeps to fixation, to try to understand how selection in diploid populations affects genetic diversity, and to learn about likely dominance mechanisms of new mutations in natural populations(Nanjundiah, 1993; Orr, 2010).

The fact that a model seen as approximating selective sweepstakes (see § 4) essentially explains population genomic data of a highly fecund population indicates that sweepstakes reproduction also facilitates rapid adaptation through selective sweepstakes. Under the Durrett-Schweinsberg model a beneficial mutation sweeps to fixation with a probability of order $O(1)$ independent of the population size, and the duration of a sweep is (on average) only of order $\log(N)$ time units(Durrett and Schweinsberg, 2005b). The variants of the Schweinsberg model(Schweinsberg, 2003) we study here, see Eq (3) and Eq (4), have yet to be compared to data; they may be shown to give as good or even better fit than the Durrett-Schweinsberg model. Extending these models of random and selective sweepstakes to the genomic scale, i.e. to several chromosomes, remains an important future task, not least since U-shaped site-frequency spectra as predicted by multiple-merger coalescents and not by the Kingman-coalescent are observed across domains of life(Freund et al., 2022). A rigorous mathematical verification of our simulation results is also an important follow-up project. However, our answer to our main question is that sweepstakes reproduction facilitates rapid adaptation where we have defined 'rapid adaptation' based on the time to fixation of a beneficial type given that it does so.

30
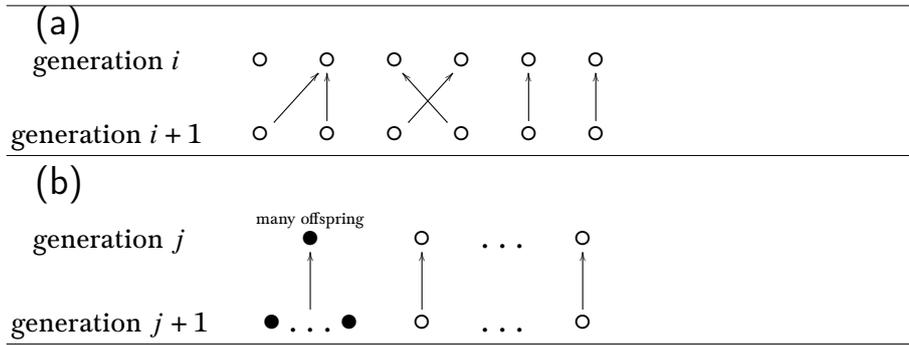
Figure 1: A simple illustration of the difference between the Wright-Fisher model (a) and a model of random sweepstakes (b), with the arrows indicating the parent of the surviving offspring. In (a) a haploid population evolves according to the Wright-Fisher model, where each individual (gene copy) produces at most a small (relative to the population size) number of surviving offspring; in (b) a single randomly picked individual contributes a large (relative to the population size) number of surviving offspring (indicated with filled circles), while all other individuals contribute the remaining number of offspring according to the Wright-Fisher model.

31

generation 0   one copy of fit type and $N-1$ copies of wild type
● ○ . . . ○

. . .

generation j two copies of fit type and $N-2$ copies of wild type
● ● ○ . . . ○

. . .

generation T_N N copies of fit type
● . . . ●

(b)

generation 0   one copy of fit type and $N-1$ copies of wild type
● ○ . . . ○

. . .

generation i m copies of fit type and $N-m$ copies of wild type
● . . . ● ○ . . . ○

generation $i+1$ $m+k$ copies of fit type and $N-m-k$ copies of wild type
● . . . ● ● . . . ● ○ . . . ○

. . .

generation T'_N N copies of fit type
● . . . ●
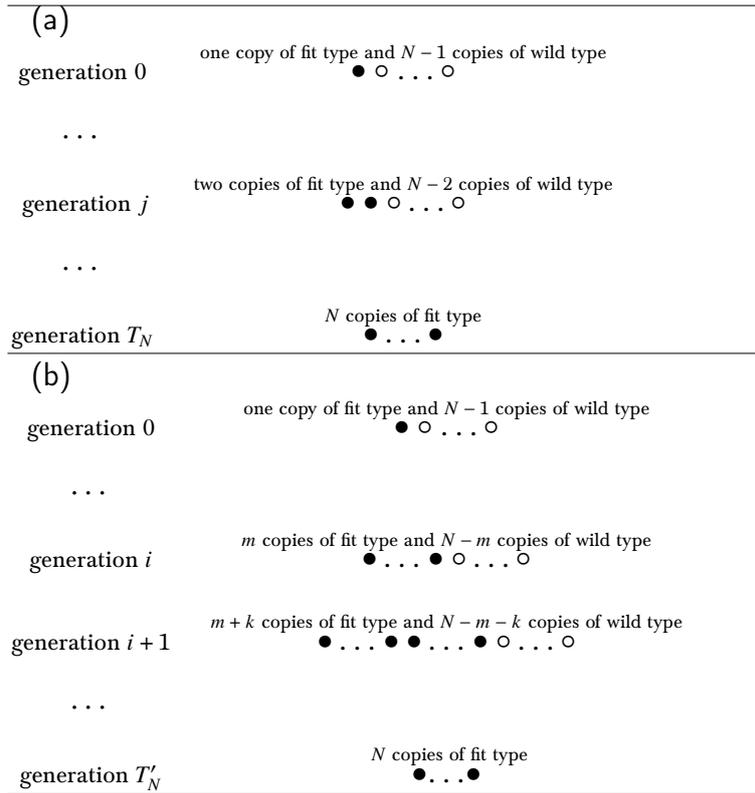
Figure 2: Random sweepstakes and fixation of an advantageous type starting in one copy. In (a) the population is assumed to evolve according to the Wright-Fisher (or similar) model, so that the number of copies of the fit type increases in small (relative to the population size) amount between generations. In (b) the population evolves according to random sweepstakes, and the number of copies of the fit type can increase by a significant amount between generations when an individual carrying the fit type produces many juveniles. On average, one would then expect, with T'_N denoting the time to fixation (conditional on fixation) under random sweepstakes and T_N under the Wright-Fisher model, $T'_N < T_N$.
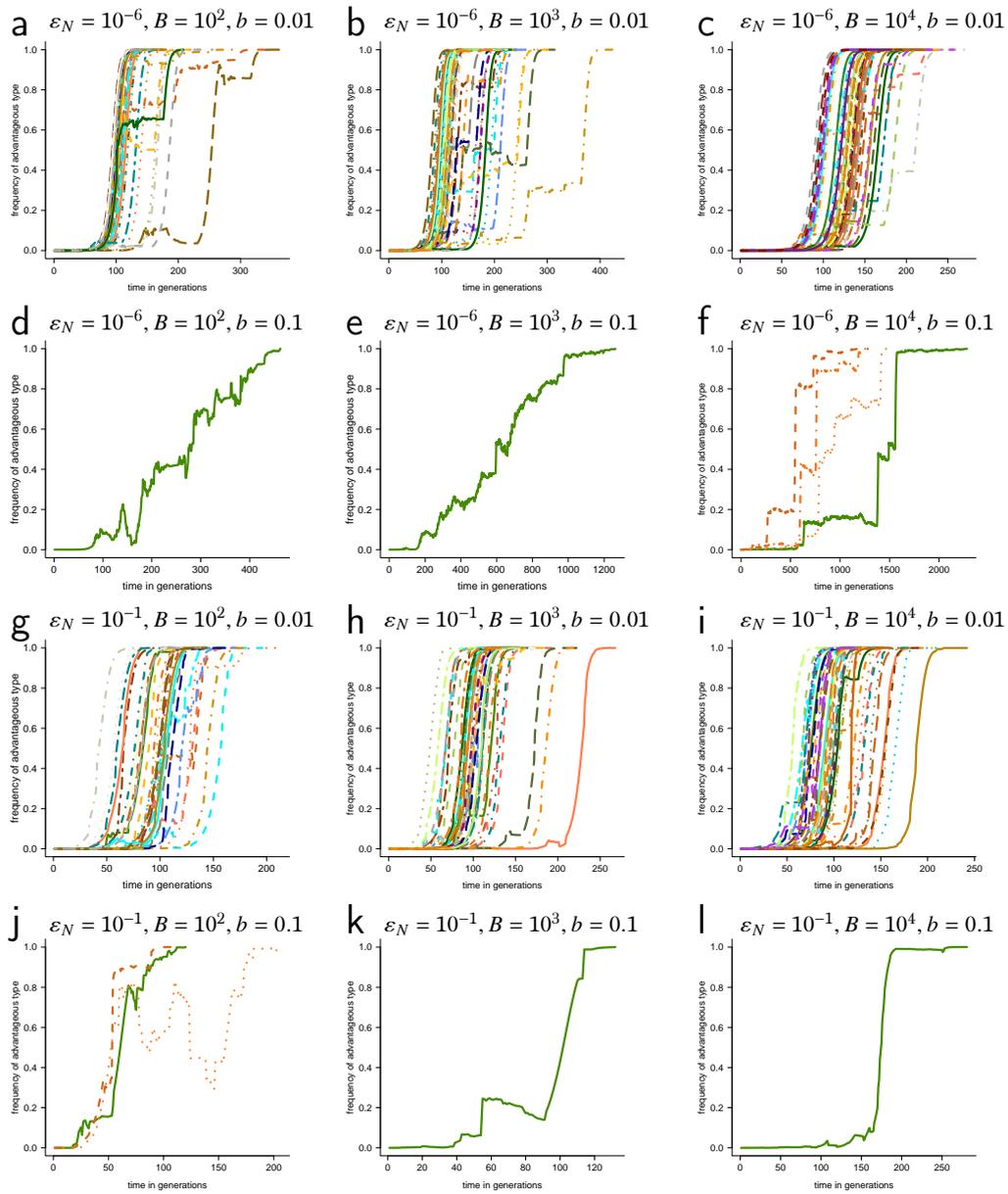
Figure 3: Random sweepstakes and recurrent bottlenecks in a haploid population. The caption is on the following page.

33

The caption for Figure 3: Examples of excursions to fixation of a type conferring a selective advantage in a haploid population of maximum size $N = 10^6$ evolving according to random sweepstakes as in Eq (4). The mechanism described in Eq (4) is that occasionally, or with probability ε_N, reproduction matches favorable conditions so that individuals have a higher chance of producing many juveniles. Most of the time however, or with probability $1 - \varepsilon_N$, individuals have a lower chance of producing many juveniles. Here we take $\alpha_1 = 0.75$ representing increased probability of producing many juveniles, $\alpha_2 = 3$ representing decreased probability of producing many juveniles. Furthermore, we take the cutoff $u(N) = N$ (each individual can produce at most $u(N)$ juveniles), strength of selection $s = 0.5$ throughout, and with ε_N the probability of having a higher chance of producing many juveniles (i.e. the probability of $\alpha = \alpha_1$), the number B of individuals surviving a bottleneck, and the probability b a bottleneck occurs in any given generation as shown. In any given generation the current individuals produce juveniles according to Eq (3) with $\alpha = \alpha_1$ with probability ε_N, and with $\alpha = \alpha_2$ with probability $1 - \varepsilon_N$, see Eq (4). Results from 10^2 (a,b,c), 16394 (j), otherwise from 10^3 experiments. The scale of the time (horizontal) axis may differ between the subplots. In each panel the trajectories, shown as the frequency of the fit type as a function of time, were obtained under identical conditions.

34
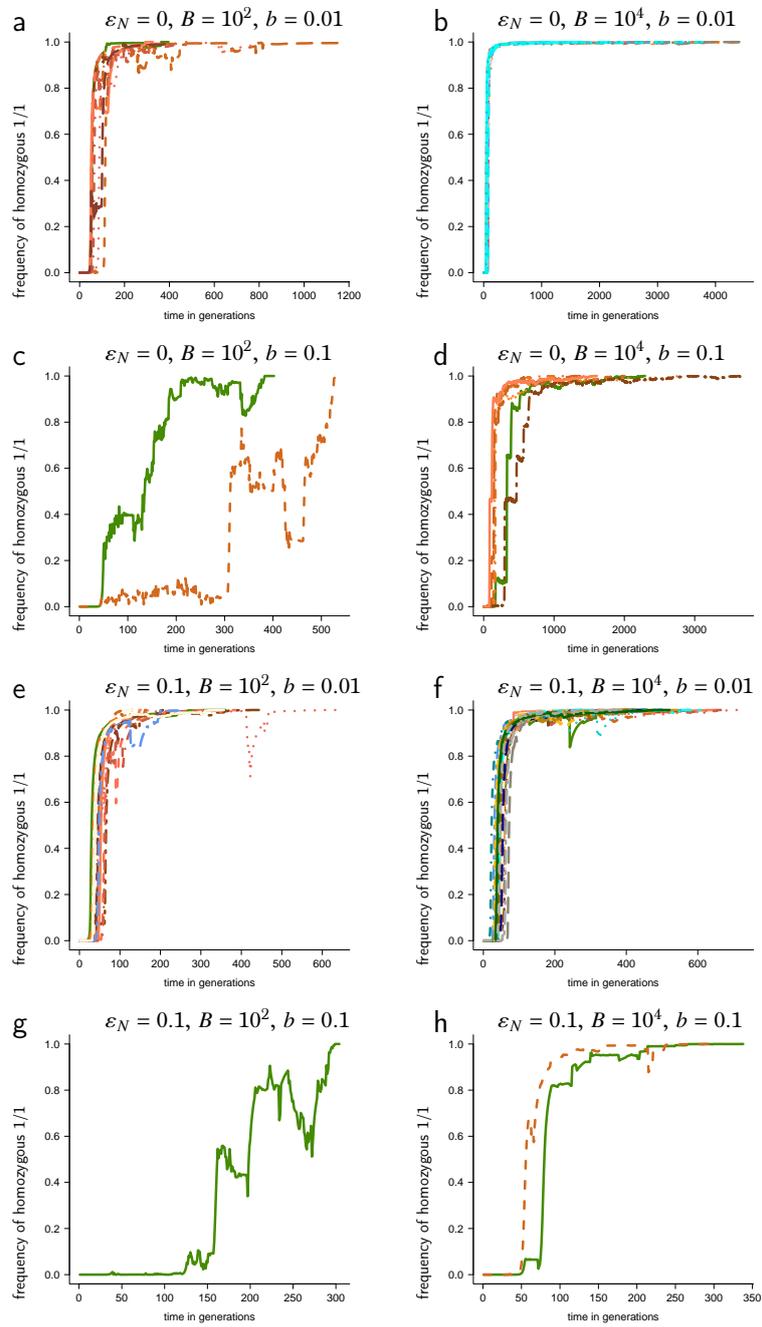
Figure 4: Complete dominance and fixation trajectories. The caption is on the following page.

Caption for Figure 4: Examples of excursions to fixation of the genotype (1/1) homozygous for the type at a single locus conferring selective advantage in a diploid population evolving according to Eq (4). The mechanism described by Eq (4) is that occasionally (with probability ε_N) individuals have a higher chance of producing many juveniles, but most of the time (with probability $1 - \varepsilon_N$) individuals produce a small number (relative to the population size) number of juveniles with high probability, and experiencing randomly occurring bottlenecks, with carrying capacity $2N = 10^6$, $\alpha_1 = 0.75$ representing a higher chance of producing many juveniles, $\alpha_2 = 3$ representing a lower chance of producing many juveniles, cutoff $u(N) = 2N$ meaning that each parent pair produces at most $u(N)$ juveniles, strength of selection $s = 0.5$ throughout, with ε_N, the number B of individuals surviving a bottleneck, and the probability b of a bottleneck in any given generation as shown. Here we consider the case of complete dominance of the beneficial type with the heterozygote as fit as the homozygote for the fit type with weight one, and the homozygote for the wild type least fit with weight e^{-4s} (see Eq S10) in § S3. Results shown from 40 experiments (a, b, e) and otherwise from 10^2 experiments. The scale of the time (horizontal) axis may differ between the subplots. In each panel the trajectories shown were obtained under identical conditions. The excursions are shown as $n_2(t)/n(t)$, where $n_2(t)$ is the number of copies of the homozygous 1/1 type at time t, and $n(t)$ is the total number of gene copies in the population at time t.
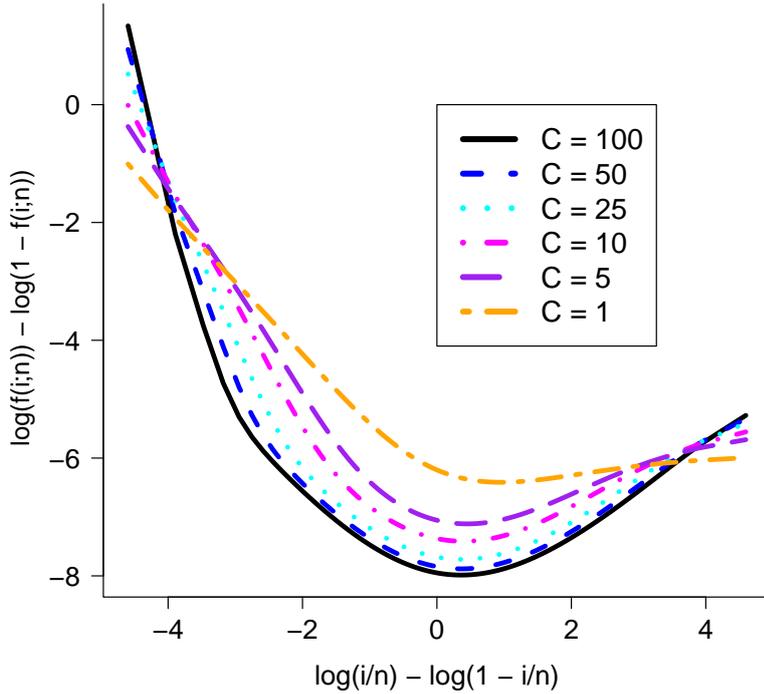
Figure 5: Recurrent sweeps predict U-shaped site-frequency spectrum. Logits (i.e. $\log(x) - \log(1 - x)$ for $0 < x < 1$) of the exact normalized expected branch length spectrum as a function of the logits of allele frequency, where $f(i;n) := \mathbb{E}\left[B_i(n)\right] / \mathbb{E}\left[B(n)\right]$, the normalized expected branch lengths supporting $i \in \{1, \ldots, n-1\}$ leaves corresponding to the size of a derived mutation (the number of times a derived mutation is observed in the sample), with sample size $n = 100$ and the parameter $\mathbb{C}$ from Eq (S6), the multiple-merger coalescent derived from the Durrett-Schweinsberg model of recurrent sweeps (Durrett and Schweinsberg, 2005a) as shown. The abscissa corresponds to a derived allele frequency (relative size of a mutation, logit scale), and the ordinate (vertical axis) corresponds to the expected number of derived mutations of a given size relative to the expected number of segregating sites (also on logit scale). The expected values were computed exactly using recursions (Birkner et al., 2013b).

References

Agrios, G. N. (2010). *Introduction to plant pathology*. Elsevier.

Árnason, E. (2004). Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885.

Árnason, E. and Halldórsdóttir, K. (2015). Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. *PeerJ*, 3:e786.

Árnason, E., Koskela, J., Halldórsdóttir, K., and Eldon, B. (2022). Sweepstakes reproductive success via pervasive and recurrent selective sweeps. *bioRxiv*.

Barfield, S., Davies, S. W., and Matz, M. V. (2022). Evidence of sweepstakes reproductive success in a broadcast-spawning coral and its implications for coral metapopulation persistence. *Molecular Ecology*.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschmar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., Rodrigues, M. F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A. W., Wong, Y., Gravel, S., Kern, A. D., Koskela, J., Ralph, P. L., and Kelleher, J. (2021). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*.

Berestycki, N. (2009). Recent progress in coalescent theory. *Ensaios Mathématicos*, 16:1–193.

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in uk biobank. *eLife*, 8:e39725.

Birkner, M. and Blath, J. (2008). Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J Math Biol*, 57:435–465.

Birkner, M. and Blath, J. (2009). Measure-valued diffusions, general coalescents and population genetic inference. In Blath, J., Mörters, P., and Scheutzow, M., editors, *Trends in stochastic analysis*, pages 329–363. Cambridge University Press.

38

Birkner, M., Blath, J., and Eldon, B. (2013a). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193:255–290.

Birkner, M., Blath, J., and Eldon, B. (2013b). Statistical properties of the site-frequency spectrum associated with Λ-coalescents. *Genetics*, 195:1037–1053.

Birkner, M., Blath, J., and Steinrücken, M. (2011). Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theor Popul Biol*, 79:155–173.

Birkner, M., Blath, J., and Steinrücken, M. (2013c). Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theor Popul Biol*, 87:15–24.

Birkner, M., Liu, H., and Sturm, A. (2018). Coalescent results for diploid exchangeable population models. *Electronic Journal of Probability*, 23(none).

Blath, J., Cronjäger, M. C., Eldon, B., and Hammer, M. (2016). The site-frequency spectrum associated with Ξ-coalescents. *Theoretical Population Biology*, 110:36–50.

Botsford, L. W., Hastings, A., and Gaines, S. D. (2001). Dependence of sustainability on the configuration of marine reserves and larval dispersal distance. *Ecology Letters*, 4(2):144–150.

Canino, M. F., Spies, I. B., Cunningham, K. M., Hauser, L., and Grant, W. S. (2010). Multiple ice-age refugia in Pacific cod, *Gadus macrocephalus. Molecular Ecology*, 19(19):4339–4351.

Cook, L. M., Grant, B. S., Saccheri, I. J., and Mallet, J. (2012). Selective bird predation on the peppered moth: the last experiment of Michael Majerus. *Biology Letters*, 8(4):609–612.

Coop, G. and Ralph, P. (2012). Patterns of neutral diversity under general models of selective sweeps. *Genetics*, 192(1):205–224.

Cowen, R. K. and Sponaugle, S. (2009). Larval dispersal and marine population connectivity. *Annual review of marine science*, 1:443–466.

Daborn, P., Yen, J., Bogwitz, M., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., et al. (2002). A single P450 allele associated with insecticide resistance in *Drosophila. Science*, 297(5590):2253–2256.

de Vladar, H. P. and Barton, N. (2014). Stability and response of polygenic traits to stabilizing selection and mutation. *Genetics*, 197(2):749–767.

Der, R., Epstein, C., and Plotkin, J. B. (2012). Dynamics of neutral and selected alleles when the offspring distribution is skewed. *Genetics*, 191(4):1331–1344.

Der, R., Epstein, C. L., and Plotkin, J. B. (2011). Generalized population models and the nature of genetic drift. *Theoretical Population Biology*, 80(2):80–99.

Desai, M. M., Walczak, A. M., and Fisher, D. S. (2013). Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585.

Donnelly, P. and Kurtz, T. G. (1999). Particle representations for measure-valued population models. *Ann Probab*, 27:166–205.

Durrett, R. and Schweinsberg, J. (2004). Approximating selective sweeps. *Theor Popul Biol*, 66:129–138.

Durrett, R. and Schweinsberg, J. (2005a). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628–1657.

Durrett, R. and Schweinsberg, J. (2005b). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch Proc Appl*, 115:1628–1657.

Eldon, B. (2011). Estimation of parameters in large offspring number models and ratios of coalescence times. *Theor Popul Biol*, 80:16–28.

Eldon, B. (2016). Inference methods for multiple merger coalescents. In Pontarotti, P., editor, *Evolutionary Biology: convergent evolution, evolution of complex traits, concepts and methods*, pages 347–371. Springer.

Eldon, B. (2020). Evolutionary genomics of high fecundity. *Annual Review of Genetics*, 54(1):213–236.

Eldon, B., Birkner, M., Blath, J., and Freund, F. (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199:841–856.

Eldon, B. and Freund, F. (2018). Genealogical properties of subsamples in highly fecund populations. *Journal of Statistical Physics*, 172(1):175–207.

Eldon, B. and Stephan, W. (2018). Evolution of highly fecund haploid populations. *Theoretical Population Biology*, 119:48–56.

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.

Etheridge, A. and Penington, S. (2022). Genealogies in bistable waves. *Electronic Journal of Probability*, 27(none):1 – 99.

Feller, W. (1951). Diffusion processes in genetics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pages 227–246. University of California Press.

Foucart, C. (2013). The impact of selection in the Λ-Wright-Fisher model. *Electron. Commun. Probab*, 18:1–10.

Freund, F., Kerdoncuff, E., Matuszewski, S., Lapierre, M., Hildebrandt, M., Jensen, J. D., Ferretti, L., Lambert, A., Sackton, T. B., and Achaz, G. (2022). Interpreting the pervasive observation of U-shaped site frequency spectra. *bioRxiv*.

Freund, F. and Siri-Jégousse, A. (2021). The impact of genetic diversity statistics on model selection between coalescents. *Computational Statistics & Data Analysis*, 156:107055.

Fu, Y.-X. and Li, W.-H. (1999). Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology*, 56(1):1–10.

Gagnaire, P.-A., Broquet, T., Aurelle, D., Viard, F., Souissi, A., Bonhomme, F., Arnaud-Haond, S., and Bierne, N. (2015). Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary applications*, 8(8):769–786.

Grant, P. R. and Grant, B. R. (2020). *How and why species multiply*. Princeton University Press.

Grant, W. S., Árnason, E., and Eldon, B. (2016). New DNA coalescent models and old population genetics software. *ICES Journal of Marine Science*, 73(9):2178–2180.

Haldane, J. (1924). A mathematical theory of natural and artificial selection. *Transactions of the Cambridge Philosophical Society*, 23(19–41).

Haller, B. C. and Messer, P. W. (2016). SLiM 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, 34(1):230–240.

Hedgecock, D. (1994). Does variance in reproductive success limit effective population sizes of marine organisms? In Beaumont, A. R., editor, *Genetics and evolution of marine organisms*, pages 122–134. Chapman and Hall, London.

Hedgecock, D. and Pudovkin, A. I. (2011). Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bull Marine Science*, 87:971–1002.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3):269–294.

Huillet, T. and Möhle, M. (2011). Population genetics models with skewed fertilities: forward and backward analysis. *Stoch Models*, 27:521–554.

Huillet, T. and Möhle, M. (2013). On the extended Moran model and its relation to coalescents with multiple collisions. *Theor Popul Biol*, 87:5–14.

Irwin, K. K., Laurent, S., Matuszewski, S., Vuilleumier, S., Ormond, L., Shim, H., Bank, C., and Jensen, J. D. (2016). On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity*, 117:393–399.

Jain, K. and Stephan, W. (2015). Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects. *G3: Genes, Genomes, Genetics*, 5(6):1065–1074.

Jain, K. and Stephan, W. (2017a). Modes of rapid polygenic adaptation. *Molecular Biology Evolution*, 34(12):3169–3175.

Jain, K. and Stephan, W. (2017b). Rapid adaptation of a polygenic trait after a sudden environmental shift. *Genetics*, 206:389–406.

John, S. and Stephan, W. (2020). Important role of genetic drift in rapid polygenic adaptation. *Ecology and Evolution*, 10(3):1278–1287.

Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989). The" hitchhiking effect" revisited. *Genetics*, 123(4):887–899.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics*, 28(4):882–901.

Kingman, J. F. C. (1982a). The coalescent. *Stoch Proc Appl*, 13:235–248.

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In Koch, G. and Spizzichino, F., editors, *Exchangeability in Probability and Statistics*, pages 97–112, Amsterdam. North-Holland.

Kingman, J. F. C. (1982c). On the genealogy of large populations. *J App Probab*, 19A:27–43.

Koskela, J. (2018). Multi-locus data distinguishes between population growth and multiple merger coalescents. *Statistical applications in genetics and molecular biology*, 17(3).

Koskela, J. and Berenguer, M. W. (2019). Robust model selection between population growth and multiple merger coalescents. *Mathematical biosciences*, 311:1–12.

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314.

Losos, J. (2009). *Lizards in an evolutionary tree*. University of California Press.

Mahrt, N., Tietze, A., Künzel, S., Franzenburg, S., Barbosa, C., Jansen, G., and Schulenburg, H. (2021). Bottleneck size and selection level reproducibly impact evolution of antibiotic resistance. *Nature ecology & evolution*, 5(9):1233–1242.

Matuszewski, S., Hildebrandt, M. E., Achaz, G., and Jensen, J. D. (2018). Coalescent processes with skewed offspring distributions and nonequilibrium demography. 208(1):323–338.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35.

Möhle, M. (2011). Coalescent processes derived from some compound Poisson population models. *Electronic Communications in Probability*, 16(none).

Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann Probab*, 29:1547–1562.

Möhle, M. and Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *J Math Biol*, 47:337–352.

Nanjundiah, V. (1993). Why are most mutations recessive? *Journal of Genetics*, 72(2-3):85–97.

Neher, R. A. and Hallatschek, O. (2013). Genealogies of rapidly adapting populations. *Proc Natl Acad Sci USA*, 110(2):437–442.

Niwa, H.-S., Nashida, K., and Yanagimoto, T. (2016). Reproductive skew in Japanese sardine inferred from DNA sequences. *ICES Journal of Marine Science: Journal du Conseil*, 73(9):2181–2189.

Orr, H. A. (2010). The population genetics of beneficial mutations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1195–1201.

Patwa, Z. and Wahl, L. M. (2008). The fixation probability of beneficial mutations. *Journal of The Royal Society Interface*, 5(28):1279–1289.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann Probab*, 27:1870–1902.

Pritchard, J. K. and Di Rienzo, A. (2010). Adaptation–not by sweeps alone. *Nature Reviews Genetics*, 11(10):665–667.

Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology*, 20(4):R208–R215.

Reznick, D. N. (2011). *The Origin then and now*. Princeton University Press.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J Appl Probab*, 36:1116–1125.

Sagitov, S. (2003). Convergence to the coalescent with simultaneous mergers. *J Appl Probab*, 40:839–854.

Sargsyan, O. and Wakeley, J. (2008). A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Pop Biol*, 74:104–114.

Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electron J Probab*, 5:1–50.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch Proc Appl*, 106:107–139.

Schweinsberg, J. (2017). Rigorous results for a population model with selection II: genealogy of the population. *Electronic Journal of Probability*, 22(none):1 − 54.

Schweinsberg, J. and Durrett, R. (2005). Random partitions approximating the coalescence of lineages during a selective sweep. *Ann Appl Probab*, 1591–1651.

Selkoe, K. A., Aloia, C. C., Crandall, E. D., Iacchei, M., Liggins, L., Puritz, J. B., von der Heyden, S., and Toonen, R. J. (2016). A decade of seascape genetics: contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, 554:1–19.

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W., Hirschhorn, J., Daly, M. J., Patterson, N., Neale, B., Mathieson, I., Reich, D., and Sunyaev, S. R. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, 8:e39702.

Stephan, W. (2016). Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular ecology*, 25(1):79–88.

Stephan, W. (2019). Selective sweeps. *Genetics*, 211(1):5–13.

Stephan, W., Wiehe, T. H., and Lenz, M. W. (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology*, 41(2):237–254.

Tajima, F. (1983). Evolutionary relationships of DNA sequences in finite populations. *Genetics*, 105:437–460.

Timm, A. and Yin, J. (2012). Kinetics of virus production from single cells. *Virology*, 424(1):11–17.

Turchin, M. C., , Chiang, C. W., Palmer, C. D., Sankararaman, S., Reich, D., and Hirschhorn, J. N. (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, 44(9):1015–1019.

van Herwaarden, O. A. and van der Wal, N. J. (2002). Extinction time and age of an allele in a large finite population. *Theoretical Population Biology*, 61(3):311–318.

van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F., and Saccheri, I. J. (2011). Industrial melanism in british peppered moths has a singular and recent mutational origin. *Science*, 332(6032):958–960.

Vendrami, D. L. J., Peck, L. S., Clark, M. S., Eldon, B., Meredith, M., and Hoffman, J. I. (2021). Sweepstake reproductive success and collective dispersal produce chaotic genetic patchiness in a broadcast spawner. *Science Advances*, 7(37).

Vignieri, S. N., Larson, J. G., and Hoekstra, H. E. (2010). The selective advantage of crypsis in mice. *Evolution: International Journal of Organic Evolution*, 64(7):2153–2158.

Wakeley, J. (2004). Recent trends in population genetics: More data! more math! simple models? *Journal of Heredity*, 95(5):397–405.

Wakeley, J. (2007). *Coalescent theory*. Roberts & Co.

Williams, G. (1975). *Sex and Evolution*. Monographs in Population Biology. Princeton University Press.

6 Data Accessibility and Benefit-Sharing

6.1 Data Accessibility Statement

The C++ code written for the simulations is freely available at
`https://github.com/eldonb/sweepstakes_reproduction_facilitates_rapid_adaptation`.
The code is also availab at Dryad with the DOI 10.5061/dryad.xwdbrv1gf

6.2 Benefit-Sharing Statement

Sweepstakes reproduction facilitates rapid adaptation in highly fecund populations

Bjarki Eldon, Wolfgang Stephan

Contents

S1   Mathematical formulation of random sweepstakes

We recall the formulation in Eq (S9). Consider a haploid (i.e. where an 'individual' corresponds to a gene copy) population of constant size N gene copies, and let $X_1, \ldots, X_N$ denote the independent and identically distributed random number of juveniles contributed by the current N individuals in any given generation. Let $\alpha > 0$ be a constant, and suppose the distribution of the random number of juveniles produced by a given individual behaves like

$$\lim_{x \to \infty} C x^{\alpha} \mathbb{P}\left(X_1 \geq x\right) = 1 \tag{S1}$$

where $C > 0$ is a constant of proportionality (i.e. ensuring that the limit in Eq (S9) is one) (Schweinsberg, 2003). The formulation in Eq (S1) should be understood as specifying how the probability of producing at least x juveniles behaves for very large x (at least on the order of the population size).

In addition to randomizing on α as in Eq (4), one can randomize on the cutoff $u(N)$ (recall Eq (3)). Suppose $u_1(N)/N \not\to 0$ (i.e. $u_1(N)$ is at least of order N), and $u_2(N)/N \to 0$

as N tends to infinity. Suppose

$$X_1, \ldots, X_N \sim \begin{cases} L(\alpha, u_2(N)) & \text{with probability } 1 - \varepsilon_N \\ L(\alpha, u_1(N)) & \text{with probability } \varepsilon_N \end{cases} \tag{S2}$$

We claim (to be verified in a planned future publication) that scalings of ε_N can be identified so that gene genealogies of a sample from a population evolving according to Eq (S2) can be described by non-trivial coalescents as population size tends to infinity.

Random sweepstakes make predictions about genetic diversity that differ significantly from predictions of the Wright-Fisher model and the Kingman-coalescent(Birkner et al., 2013; Blath et al., 2016); for example multiple-merger coalescents derived from models of random sweepstakes can be distinguished from the extended (time-changed) Kingman-coalescent incorporating population growth(Eldon et al., 2015; Koskela, 2018; Koskela and Berenguer, 2019). The mathematical formulation of random sweepstakes is reviewed in (Birkner and Blath, 2009). Briefly, and without going into technical details, the process $\{Y_t; t \geq 0\}$ tracking the frequency going forward in time of a given genetic type in a population evolving according to random sweepstakes is the solution of the stochastic differential equation (where Y_{t-} denotes the state of the process just before time t)

$$dY_t = \sqrt{Y_{t-}(1 - Y_{t-})}dW_t + \text{jumps} \tag{S3}$$

In Eq (S3) $W_t \equiv \{W_t; t \geq 0\}$ denotes standard Brownian motion, and 'jumps' refer to a stochastic process governing the discontinuous changes in the type frequency generated by random sweepstakes (Birkner and Blath, 2009). Eq (S3) is a key formulation of random sweepstakes and in a nutshell shows why models of random sweepstakes are essential for understanding genetic diversity in highly fecund populations. A type frequency in a population evolving according to the Wright-Fisher model is a Wright-Fisher diffusion, or the (unique) solution of the SDE

$$dY_t = \sqrt{Y_{t-}(1 - Y_{t-})}dW_t \tag{S4}$$

Equation (S3) describes the evolution of a type frequency in a population evolving according to a model similar to the one in Eq (S8). The coalescent corresponding to Eq (S3) is a (multiple-merger) Λ-coalescent (Donnelly and Kurtz, 1999; Pitman, 1999; Sagitov, 1999) with Λ-measure of the form $\Lambda = \delta_0 + \Lambda_0$, where Λ_0 is a finite measure without an atom at zero. The corresponding rate at which a given group of k out of $n \geq 2$ lineages merge is

$$\lambda_{n,k} = \mathbb{1}_{\{k=2\}} + \int_0^1 x^{k-2}(1 - x)^{n-k}\Lambda(dx) \tag{S5}$$

The coalescent in Eq (S5) is a mixture of the Kingman-coalescent (the '$\mathbb{1}_{\{k=2\}}$' part, and a multiple-merger coalescent. The Durrett-Schweinsberg model of recurrent sweeps leading to a multiple-merger coalescent (see Eq (S6)) is of this form(Durrett and Schweinsberg, 2005a), the model of random sweepstakes investigated in (Eldon and Wakeley, 2006) (see Eq 5) can also lead to a coalescent of the form in Eq (S5), and more general versions of Eq (5) investigated by Huillet and Möhle (2013). The rate at which a given group of k out of $n \geq 2$ lineages merge in the multiple-merger coalescent derived from the Durrett-

Schweinsberg model of recurrent sweeps is given by

$$\lambda_{n,k} = \mathbb{1}_{\{k=2\}} + 2\mathbb{C}B(k, n-k+1), \quad 2 \le k \le n \tag{S6}$$

(see Example 2.4 inDurrett and Schweinsberg (2005a)). In (S6) $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function, and $\mathbb{C} > 0$ is a composite parameter reflecting the strength of mutation, selection, and recombination (Durrett and Schweinsberg, 2004, 2005b). Values of $\mathbb{C}$ between six and ten roughly give the best fit of the model to the Atlantic cod data(Árnason et al., 2022).

To further understand the effect of random sweepstakes on selection Der et al. (2011) propose a generalised Wright-Fisher model. Let $\{Y_t\}$ denote the number of copies of a given type at time t in a haploid population of size N and suppose (Der et al., 2011)

$$\mathbb{E}\left[Y_{t+1}|Y_t\right] = Y_t$$
$$\text{Var}\left[Y_{t+1}|Y_t\right] = \frac{N\text{Var}[Y_{t+1}|\{Y_t=1\}]}{N-1}Y_t\left(1 - \frac{Y_t}{N}\right) \tag{S7}$$

i.e. $\{Y_t\}$ is restricted to being a martingale. Der et al. (2011) derive bounds on the fixation probability and time to absorption (i.e. of a type either fixing or being lost from a population) for processes described by Eq (S7). For comparison with the model of evolution described by Eq (S7), the transition probability for a type frequency in our framework (Eq (3)) in the haploid case, in the absence of selection and bottlenecks, is

$$\mathbb{P}\left(Y_{t+1} = y|Y_t = j\right) = \mathbb{E}\left[\frac{\binom{S_j}{y}\binom{S_{N-j}}{N-y}}{\binom{S_j+S_{N-j}}{N}}\right] \tag{S8}$$

in a haploid population of fixed size N partitioned into two types where S_j is the random number of juveniles produced by j individuals of the type we are tracking, and S_{N-j} denotes the random number of juveniles produced by the $N-j$ individuals of the other type. Applying continuum techniques would require identifying an operator G acting on a test function u with

$$Gu = \lim_{N\to\infty} N(P_N - I)u_N \tag{S9}$$

(Der et al., 2011) where P_N is the transition probability matrix corresponding to Eq (S8). Our model in Eq (3) resembles the power-law processes studied by Der et al. (2011).

S2 Examples of trajectories for a haploid population

Write $T_k(y) := \min\{t \geq 0 : Y_t = k, Y_0 = y\}$ for the first time we see k copies of the beneficial type when starting with y copies. Then we consider the quantities

$$
\begin{aligned}
p_N(1) &:= \mathbb{P}\left(T_N(1) < T_0(1)\right), \\
\tau_N(1) &:= \mathbb{E}\left[T_N(1)|T_N(1) < T_0(1)\right].
\end{aligned}
\tag{S10}
$$

In this section we give examples of trajectories to fixation of an advantageous type in a haploid population of constant size evolving according to Eq (4). The biological interpretation of Eq (4) is that most of the time (with probability $1 - \varepsilon_N$) individuals produce juveniles with a small probability of producing many ($\alpha = \alpha_2$ in Eq (3)), but once in a while (with probability ε_N) reproduction matches favorable conditions so that individuals produce juveniles with a higher probability of producing many juveniles ($\alpha = \alpha_1$).

Recall that ε_N is the probability of individuals producing juveniles according to Eq (3) with $\alpha = \alpha_1$, i.e. with a higher probability of producing many juveniles. The case $\varepsilon_N = 0$ corresponds to the absence of random sweepstakes, i.e. only small (relative to the population size) families occur with high probability since then $\alpha = \alpha_2 = 3$; the case $\varepsilon_N > 0$ corresponds to the presence of random sweepstakes ($\alpha = \alpha_1$ in Eq (3)); the value $\varepsilon_N = 0.1$ as well as the value $\alpha_1 = 0.75$ were chosen to give an example of 'strong' effect of random sweepstakes; the smaller the value of α in Eq (3) the higher the chance of producing many juveniles; with $u(N) = N$, i.e. each individual can produce at most N juveniles.

S2.1 Random sweepstakes without bottlenecks

In this section we give examples of trajectories to fixation for a haploid population of constant size evolving according to Eq (4) without recurrent bottlenecks. The interpretation from the particular scenario considered in Figure S1 is that the effect of random sweepstakes on the time to fixation ($\tau_N(1)$) is negligible.
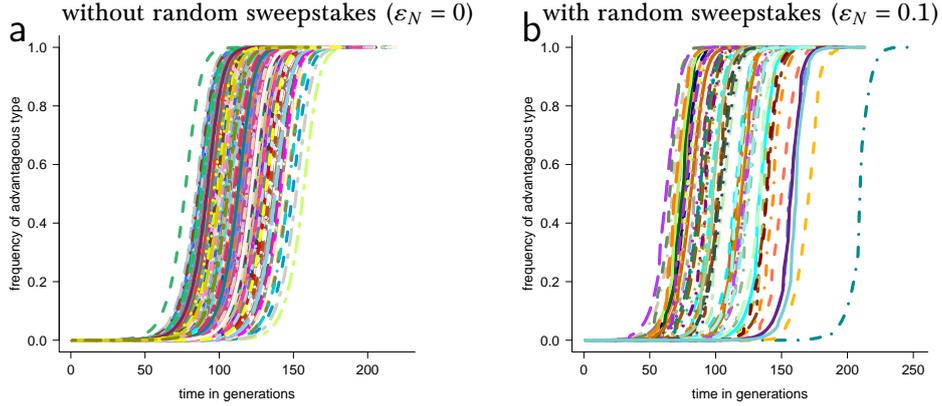
Figure S1: Random sweepstakes without bottlenecks. Examples of excursions to fixation of the type conferring a selective advantage in a haploid population of fixed size $N = 10^6$ and evolving according to Eq (4) with $\varepsilon_N = 0$ (a; absence of random sweepstakes) and 0.1 (b; presence of random sweepstakes) with cutoff $u(N) = N$ i.e. individuals can produce at most N juveniles (see Eq (3)), $\alpha_1 = 0.75$ representing increased probability of producing many juveniles, $\alpha_2 = 3$ (representing decreased probability of producing many juveniles), and strength of selection $s = 0.5$, i.e. juveniles with the wild type have viability weight e^{-s}, juveniles with the beneficial type have viability weight one, with 606 trajectories (a) and 61 (b) out of 10^3 experiments. Equation (4) means that in any given generation the current individuals produce juveniles according to Eq (3) with $\alpha = \alpha_1$ with probability ε_N, and according to $\alpha = \alpha_2$ with probability $1 - \varepsilon_N$; thus in (a) $\alpha = \alpha_2$ all the time since $\varepsilon_N = 0$. In each experiment we start with the beneficial type in one copy. The excursions are shown as $n_1(t)/n(t)$, where $n_1(t)$ is the number of copies of the beneficial type at time t, and $n(t)$ is the total number of gene copies at time t. In each panel the different line types and colours are only meant to distinguish between the trajectories; in (a) all the trajectories were obtained under identical conditions, and the same goes for (b).

S2.2 Recurrent bottlenecks without large families

Figure S2 shows examples of excursions for a haploid population evolving according to Eq (3) with $\alpha = 3$, i.e. individuals produce small families with high probability (so the evolution of the population is effectively the same is if it was evolving according to the Wright-Fisher model), and bottlenecks of increasing size from left to right occur randomly with probability 0.01 in any given generation in the top row (showing the excursions from 10^2 experiments), and 0.1 in the bottom row (showing the excursions from 10^4 experiments). The effects of randomly occurring bottlenecks on $p_N(1)$ (the probability of fixation of the beneficial type when starting with the type in one copy, Eq (S10)) and $\tau_N(1)$ (the average time to fixation conditional on fixation, Eq (S10)) are evident. If the probability of a bottleneck is 'small' (0.01; top row in Fig S2) varying the size of the bottleneck does little to change $\tau_N(1)$. High probability of a bottleneck (0.1; bottom row in Fig S2) reduces $p_N(1)$, and if the bottleneck is not too severe it can drastically increase $\tau_N(1)$.
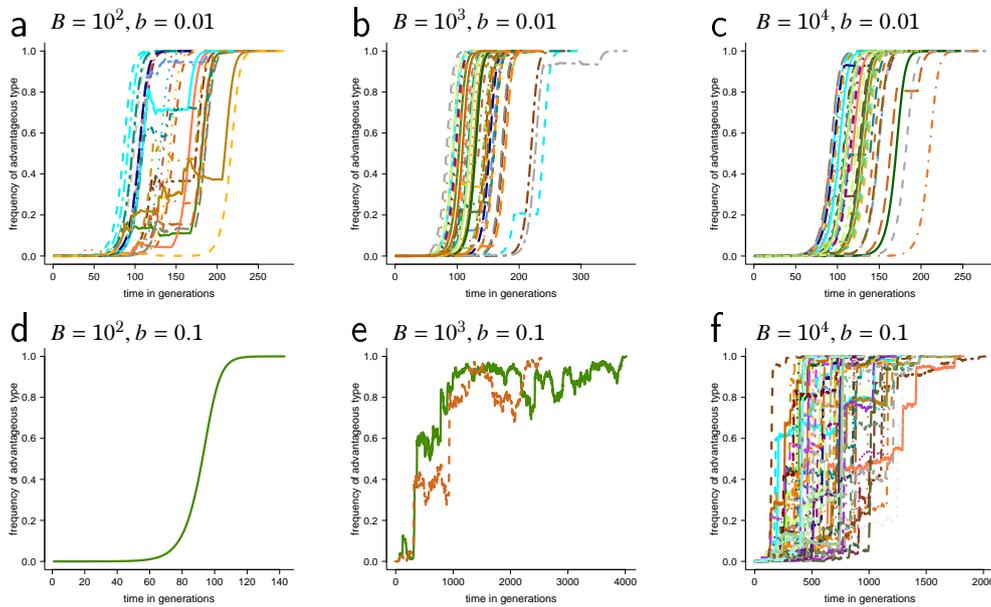

Figure S2: Recurrent bottlenecks without large families. Examples of excursions to fixation of a type conferring a selective advantage in a haploid population of maximum size $N = 10^6$ evolving according to Eq (3) with $\alpha = 3$ (i.e. each individual produces a small number, relative to the population size, of juveniles with high probability), cutoff $u(N) = N$, and strength of selection $s = 0.5$ throughout; bottlenecks of size B as shown and with probability of a bottleneck in any given generation (b) as shown. Results from 10^2 (a,b,c) experiments, otherwise from 10^4 experiments. The scale of the abscissa (time axis) may differ between the subplots. In each panel the excursions shown were all obtained under identical conditions.

S3 Diploid populations and dominance mechanisms

In this section we investigate the effect of various dominance mechanisms on selection in a diploid population evolving according to Eq (4), where most of the time (with probability $1 - \varepsilon_N$) the current parent pairs independently produce a small number (relative to the population size) of juveniles with high probability (since then α in Eq (3) is increased); occasionally reproduction matches favorable environmental conditions and the current parent pairs have a higher chance of producing a larger number of juveniles (since then α in Eq (3) is decreased).

The genotype of each given juvenile determines the viability weight according to

$$w = \exp\left(-s\left(z_0 - z(g)\right)^2\right) \tag{S11}$$

where z_0 is the optimal trait value, $z(g)$ is the trait value of a juvenile with genotype g, and $s \geq 0$ the strength of selection.

Let g in Eq (S11) take values $\{0, 1, 2\}$, where 0 denotes homozygous wild type 0/0, 1 denotes heterozygous type 0/1, and 2 denotes type 1/1 homozygous for the type conferring selective advantage. Taking $z_0 = 2$ in Eq (S11), the cases we will consider are, with dominance mechanism relative to the type conferring selective advantage,

$$\begin{cases} \text{complete dominance} & 0 = z(0) < z(1) = z(2) = 2, \\ \text{incomplete dominance} & 0 = z(0) < z(1) < z(2) = 2, \\ \text{recessive} & 0 = z(0) = z(1) < z(2) = 2. \end{cases} \tag{S12}$$

S3.1 Random sweepstakes without bottlenecks

In this section we investigate the effect of random sweepstakes on selection in a diploid population of constant size evolving according to Eq (4). In Figure S3 we compare the joint effects of dominance mechanism and random sweepstakes (no bottlenecks) for a diploid population evolving according to Eq (4) with $\varepsilon_N = 0$ (left column) and 0.1 (right column) with $\alpha_1 = 0.75$ and cutoff $u(N) = 2N = 10^6$, the number of diploid individuals. The dominance mechanisms relative to the beneficial type (Eq (S12)) are incomplete (top row), complete (middle row), and recessive (bottom row). The number of excursions in the right column ($\varepsilon_N = 0.1$) is roughly half the number in the left column ($\varepsilon_N = 0$) for the same number of experiments; random sweepstakes reduce $p_N(1)$ but not by much. The effect of random sweepstakes on the time to fixation depends on the dominance mechanism; if the dominance mechanism is incomplete the effect is hardly noticeable, but the story is quite different for the other two dominance mechanisms. The time to fixation (conditional on fixation) is clearly increased in both the complete dominance (middle row), and recessive beneficial type (bottom row) in the absence of random sweepstakes, and random sweepstakes act to reduce the time to fixation (compare the horizontal scale of the graphs). Furthermore, $p_N(1)$ is clearly reduced when the beneficial type is recessive (the number of experiments is 10^4 for the bottom row, and 10^2 otherwise).
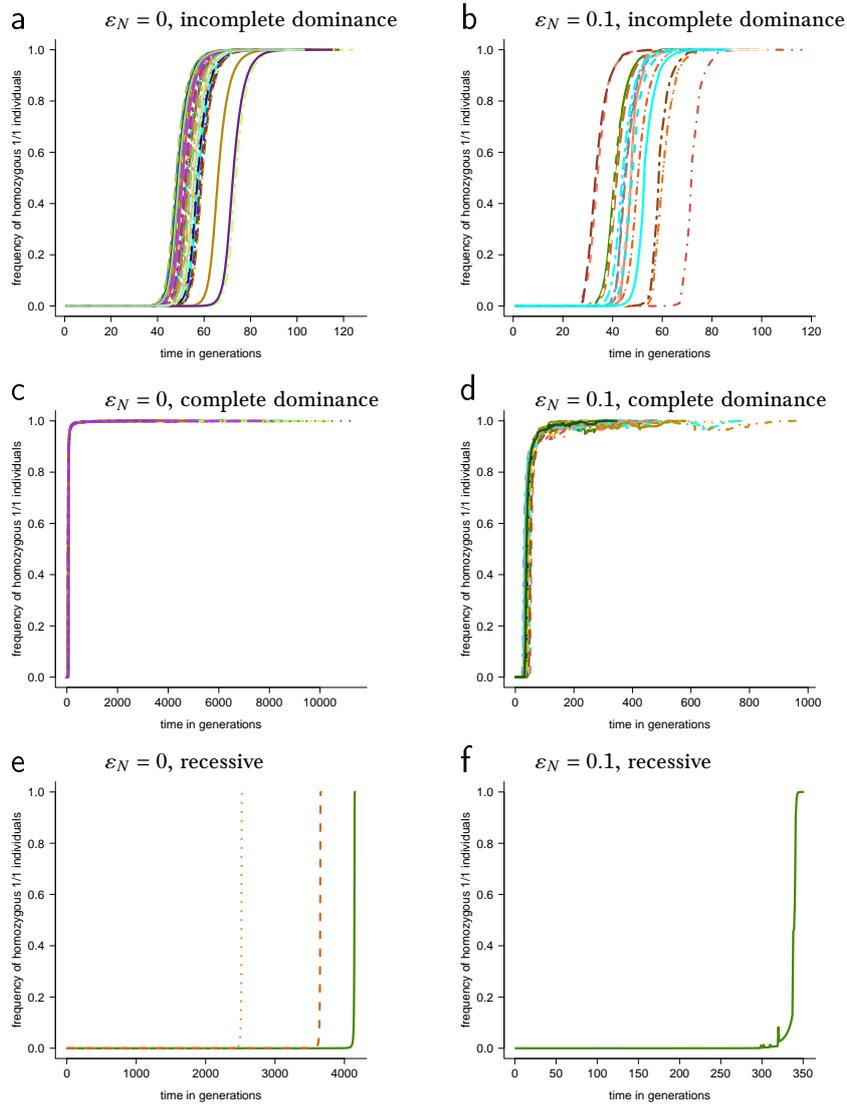
Figure S3: Random sweepstakes without bottlenecks. Examples of excursions to fixation of the genotype (1/1) homozygous for the type at a single locus conferring selective advantage in a diploid population of constant size $2N = 10^6$ diploid individuals evolving according to Eq (4), i.e. most of the time, or with probability $1-\varepsilon_N$, the current parent pairs produce a small number (relative to the population size) number of juveniles with high probability since then α in Eq (3) is increased ($\alpha = \alpha_2$); occasionally, i.e. with probability ε_N, reproduction matches favorable conditions and the parent pairs produce juveniles according to Eq (3) with α in Eq (3) decreased ($\alpha = \alpha_1$). Throughout we have $\alpha_1 = 0.75$, $\alpha_2 = 3$, cutoff $u(N) = 2N$ (each parent pair can produce at most $u(N)$ juveniles), strength of selection $s = 0.5$ (recall Eq (S11)), and with ε_N and dominance mechanisms (Eq (S12)) as shown. The trajectories, shown as the number of diploid individuals homozygous for the fit type relative to the population size as a function of time were obtained for 10^2 experiments (a,b,c,d), otherwise 10^4 experiments (e,f). The scale of the abscissa (time axis) may differ between the subplots. In each panel the trajectories shown were obtained under identical conditions.

8

S3.2 Incomplete dominance of the fit type

In this section we investigate the case when the beneficial type in a diploid population evolving according to random sweepstakes (Eq (4)) and recurrent bottlenecks when the beneficial type is given incomplete dominance (Eq (S12)). In the case of incomplete dominance (Figure S4) a noticeable effect of random sweepstakes (bottom two rows, $\varepsilon_N = 0.1$) is to reduce $\tau_N(1)$ (Eq (S10)), while bottlenecks (bottlenecks occur with probability 0.01 per generation in first and third row, and with probability 0.1 in second and fourth row) tend to increase $\tau_N(1)$.
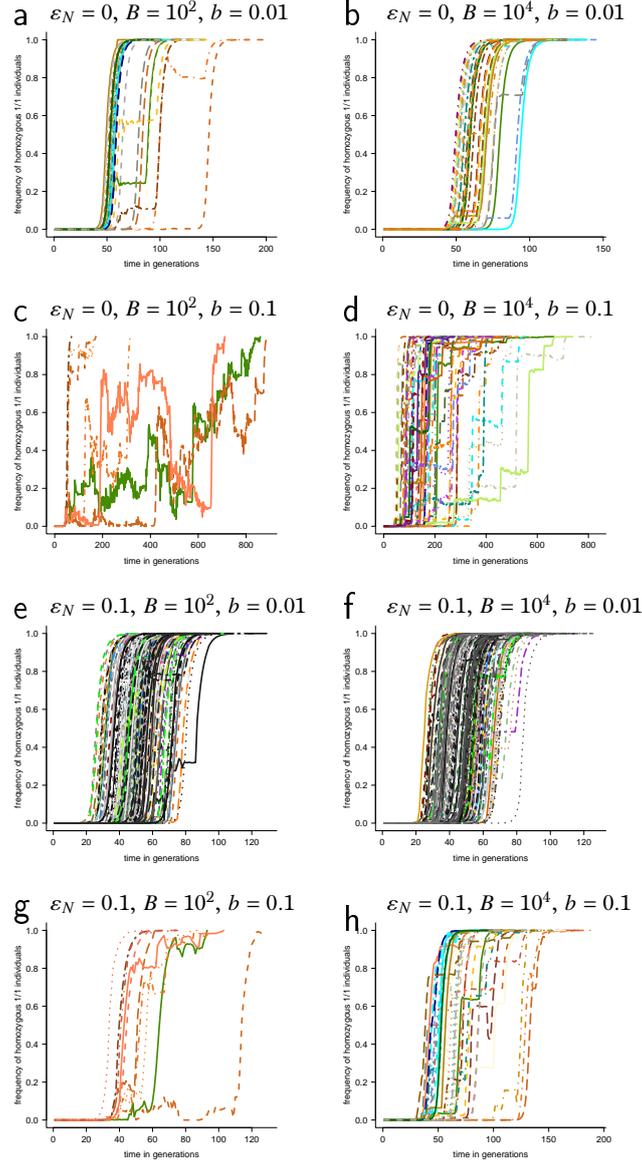
Figure S4: Incomplete dominance of the fit type. Examples of excursions to fixation of the genotype (1/1) homozygous for the type at a single locus conferring selective advantage in a diploid population evolving according to Eq (4). The evolution described by Eq (4) is that most of the time, or with probability $1 - \varepsilon_N$, the current parent pairs produce a small number (relative to the population size) number of juveniles with high probability since then α in Eq (3) is increased ($\alpha = \alpha_2$); occasionally, i.e. with probability ε_N, reproduction matches favorable conditions and the parent pairs produce juveniles according to Eq (3) with α in Eq (3) decreased ($\alpha = \alpha_1$). Furthermore, the population experiences randomly occurring bottlenecks, with carrying capacity $2N = 10^6$, $\alpha_1 = 0.75$, $\alpha_2 = 3$, cutoff $u(N) = 2N$, strength of selection $s = 0.5$ throughout, with ε_N, bottleneck size B, and the probability b of a bottleneck in any given generation as shown. Here we consider the case of incomplete dominance of the beneficial type (Eq (S12)). The trajectories, shown as the number of diploid individuals homozygous for the fit type relative to the population size as a function of time were obtained for 10^2 experiments (a,b), otherwise from 10^3 experiments. The scale of the abscissa (time axis) may differ between panels. In each panel the excursions shown were all obtained under identical conditions.

10

S3.3 The fit type as recessive

In this section investigate the effect of random sweepstakes and recurrent bottlenecks on selection at a single site in a diploid population paritioned into two types with the fit type as recessive (Fig S5). The effect of random sweepstakes ($\varepsilon_N > 0$) on the time to fixation (conditional on fixation), especially in the presence of frequently occurring bottlenecks, is clear.

a $\varepsilon_N = 0$, $B = 10^2$, $b = 0.01$

b $\varepsilon_N = 0$, $B = 10^4$, $b = 0.01$

c $\varepsilon_N = 0$, $B = 10^2$, $b = 0.1$

d $\varepsilon_N = 0$, $B = 10^4$, $b = 0.1$

e $\varepsilon_N = 0.1$, $B = 10^2$, $b = 0.01$

f $\varepsilon_N = 0.1$, $B = 10^4$, $b = 0.01$

g $\varepsilon_N = 0.1$, $B = 10^2$, $b = 0.1$

h $\varepsilon_N = 0.1$, $B = 10^4$, $b = 0.1$
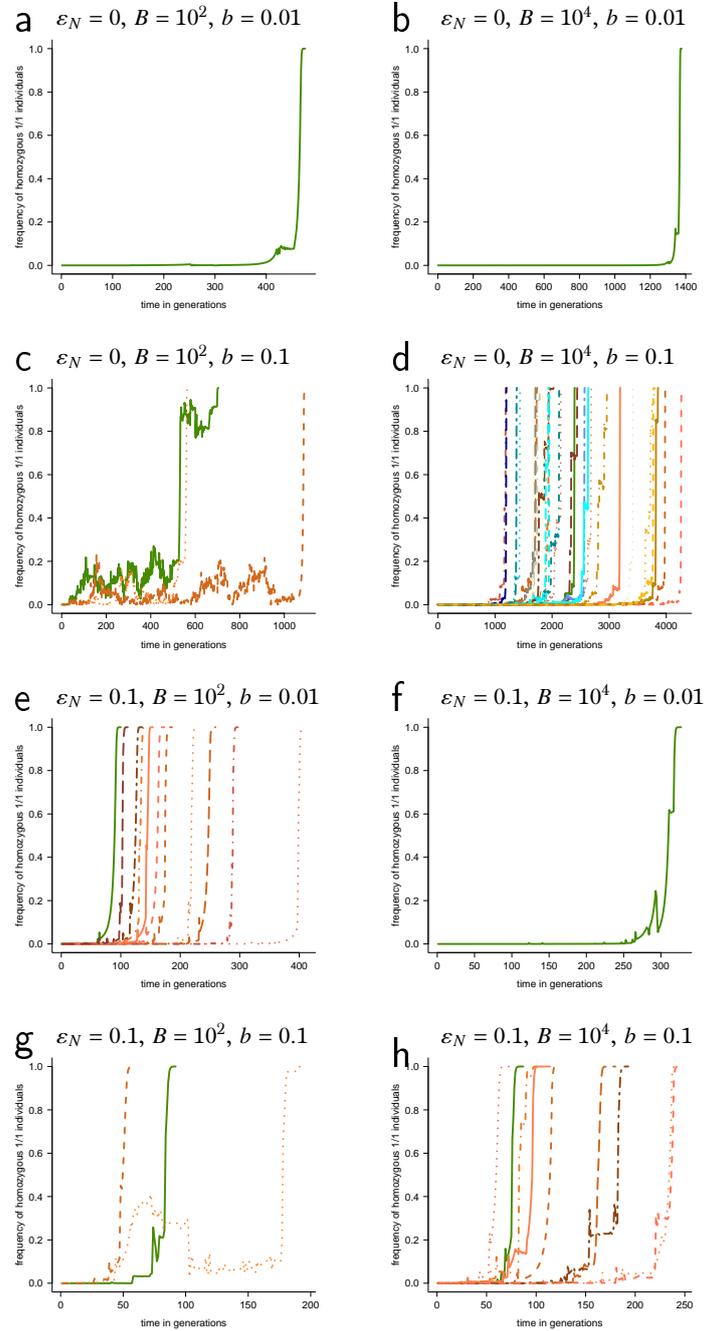
Figure S5: A recessive fit type. The caption is on the following page.

12

Caption to Figure S5: The fit type as recessive. Examples of excursions to fixation of the genotype 1/1 homozygous for the type at a single locus conferring selective advantage in a diploid population evolving according to Eq (4) where occasionally (with probability ε_N) reproduction matches favorable conditions so that individuals have a higher chance of producing many juveniles, but most of the time (with probability $1 - \varepsilon_N$) individuals produce a small number (relative to the population size) number of juveniles with high probability, and experiencing randomly occurring bottlenecks, with carrying capacity $2N = 10^6$, $\alpha_1 = 0.75$ representing an higher chance of producing many juveniles , $\alpha_2 = 3$ representing a lower chance of producing many juveniles, cutoff $u(N) = 2N$ meaning that each parent pair can produce at most $u(N)$ juveniles, strength of selection $s = 0.5$ throughout, with ε_N, the number B of individuals surviving a bottleneck, and the probability b of a bottleneck in a given generation as shown. The dominance mechanism is with the beneficial type recessive (Eq (S12)). Results shown from 10^4 (a,b,f) otherwise from 10^6 experiments. The excursions are shown as $n_2(t)/n(t)$ as a function of time t in generations, where $n_2(t)$ is the number of copies of the homozygous 1/1 type at time t, and $n(t)$ is the total number of gene copies in the population at time (generation) t. The scale of the abscissa (time axis) may vary between panels. In each panel the trajectories were obtained under identical conditions.

S4   Illustration of the Durrett-Schweinsberg model

In this section the Durrett-Schweinsberg model(Durrett and Schweinsberg, 2005a) of a fixation of a beneficial type arising by mutation in a population evolving according to the Moran model (Figure S6). The mutation occurs at a site linked to a 'neutral' site, i.e. a site assumed to never experiencing mutation. During a sweep recombination between the neutral site and the site experiencing mutation can lead to some ancestral lineages of a sample to 'escape' a sweep by moving onto a different background. If that happens, not all ancestral lineages merge during a sweep, leading to a non-trivial multiple merger genealogy of the neutral site (Eq (S6)).

## The Durrett-Schweinsberg model of a sweep
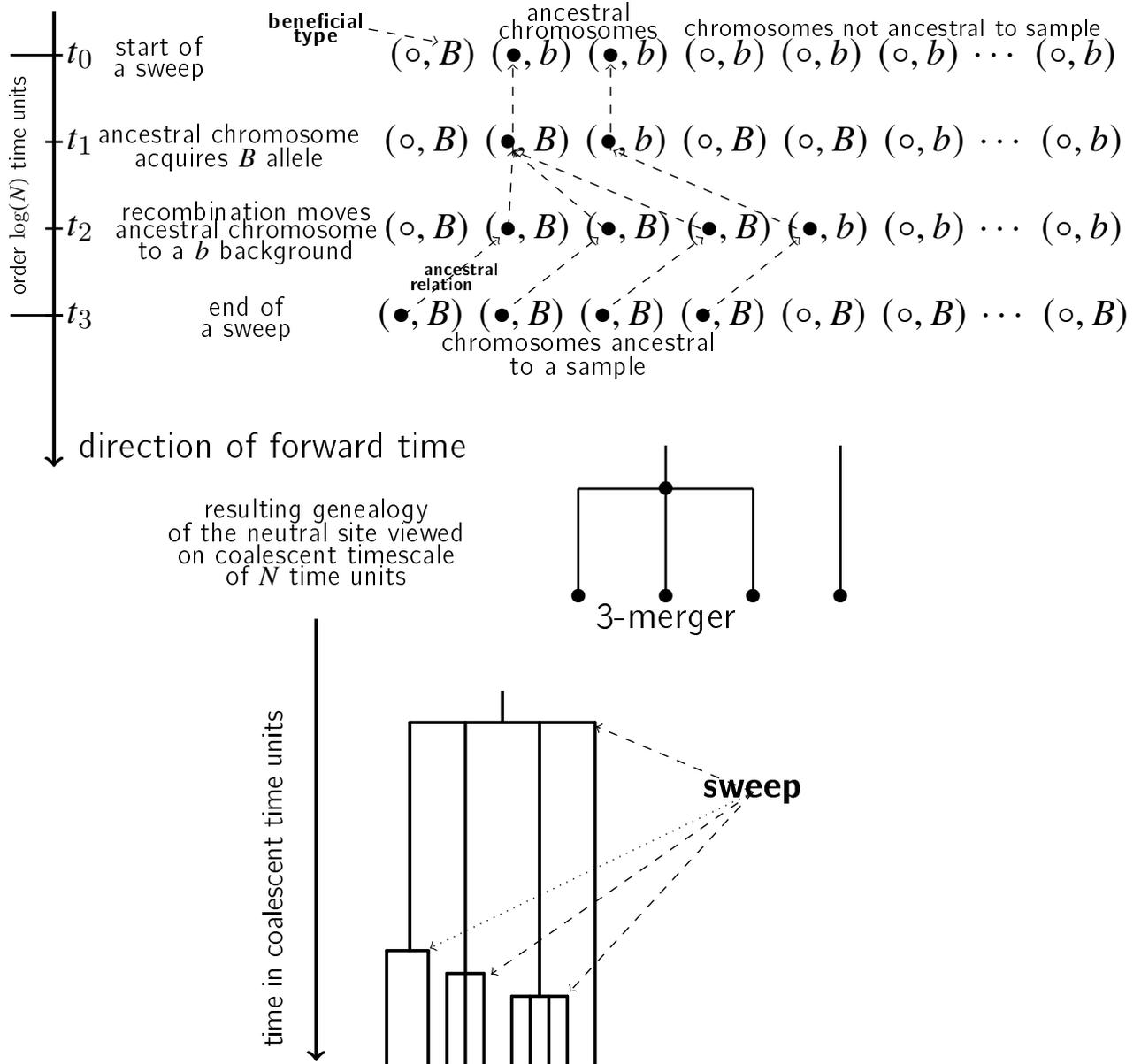
Figure S6: Illustration of the Durrett-Schweinsberg model (Durrett and Schweinsberg, 2005a) of a sweep leading to a multiple merger in the genealogy at the neutral site. The beneficial type is denoted B, and arises through a mutation. The gene copies of the neutral site are denoted as empty circles or filled circled when ancestral to a sample. During the sweep, one ancestral chromosome escapes a sweep through recombination. A sweep takes on the order of $\log(N)$ time units, so occurs instantaneously when time is viewed in N time units, leading to an apparent multiple merger in the genealogy of the neutral site. In the example genealogy at the bottom, pairwise mergers may be due to a sweep.

S5   The expected number of segregating sites

For completeness we state in Eq (S13) the expected number of segregating sites (assuming the infinitely-many sites mutation model) for a sample of size n gene copies. Each gene copy represents a contiguous non-recombining segment of a chromosome of lengths L base pairs, with μ denoting the per site per generation mutation rate. We denote by $B(n)$ the random total size of a genealogy of a sample of size n, and $M(n)$ denotes the random number of segregating sites (derived mutations) in a sample of size n. Then

$$\frac{1}{c_N}\mu L \mathbb{E}\left[B(n)\right] = \mathbb{E}[M(n)] \tag{S13}$$

A simple rearrangement yields Eq (2).

S6   Describing a fixation trajectory

For completeness we state in this section, see Eq (S14), the logistic differential equation that has been used to describe the trajectory of a mutation to fixation conditional on the mutation doing so,

$$\frac{dp_t}{dt} = sp_t(1-p_t) \tag{S14}$$

In Eq (S14), p_t denotes the frequency of the beneficial type at time t, and $s > 0$ the strength of selection(Kaplan et al., 1989; Stephan et al., 1992). Schweinsberg and Durrett (2005) suggest another approach for describing fixation trajectories.

S7   Literature cited

References

Árnason, E., Koskela, J., Halldórsdóttir, K., and Eldon, B. (2022). Sweepstakes reproductive success via pervasive and recurrent selective sweeps.

Birkner, M. and Blath, J. (2009). Measure-valued diffusions, general coalescents and population genetic inference. In Blath, J., Mörters, P., and Scheutzow, M., editors, *Trends in stochastic analysis*, pages 329–363. Cambridge University Press.

Birkner, M., Blath, J., and Eldon, B. (2013). Statistical properties of the site-frequency spectrum associated with Λ-coalescents. *Genetics*, 195:1037–1053.

Blath, J., Cronjäger, M. C., Eldon, B., and Hammer, M. (2016). The site-frequency spectrum associated with Ξ-coalescents. *Theoretical Population Biology*, 110:36–50.

Der, R., Epstein, C. L., and Plotkin, J. B. (2011). Generalized population models and the nature of genetic drift. *Theoretical Population Biology*, 80(2):80–99.

Donnelly, P. and Kurtz, T. G. (1999). Particle representations for measure-valued population models. *Ann Probab*, 27:166–205.

Durrett, R. and Schweinsberg, J. (2004). Approximating selective sweeps. *Theor Popul Biol*, 66:129–138.

Durrett, R. and Schweinsberg, J. (2005a). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628–1657.

Durrett, R. and Schweinsberg, J. (2005b). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch Proc Appl*, 115:1628–1657.

Eldon, B., Birkner, M., Blath, J., and Freund, F. (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199:841–856.

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.

Huillet, T. and Möhle, M. (2013). On the extended Moran model and its relation to coalescents with multiple collisions. *Theor Popul Biol*, 87:5–14.

Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989). The" hitchhiking effect" revisited. *Genetics*, 123(4):887–899.

Koskela, J. (2018). Multi-locus data distinguishes between population growth and multiple merger coalescents. *Statistical applications in genetics and molecular biology*, 17(3).

Koskela, J. and Berenguer, M. W. (2019). Robust model selection between population growth and multiple merger coalescents. *Mathematical biosciences*, 311:1–12.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann Probab*, 27:1870–1902.

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J Appl Probab*, 36:1116–1125.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch Proc Appl*, 106:107–139.

Schweinsberg, J. and Durrett, R. (2005). Random partitions approximating the coalescence of lineages during a selective sweep. *Ann Appl Probab*, 1591–1651.

Stephan, W., Wiehe, T. H., and Lenz, M. W. (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology*, 41(2):237–254.