Dear Editor and Reviewers,

Thank you for the many thoughtful comments. We addressed them in the current version of the manuscript and highlighted our responses below.

Kind regards,
Kevin Korfmann

by Renaud Vitalis, 06 Apr 2023 18:36
Manuscript: **https://doi.org/10.1101/2022.04.26.489499** version 2

Dear Dr Korfmann,

I apologize for the delay. Three reviewers have now examined the revised version of your manuscript entitled "Weak seed banks influence the signature and detectability of selective sweeps" that you submitted for recommendation to PCI Evol Biol.

They all reckon that you made a very substantial effort to address the issues raised in the previous evaluation round. I agree with them that your revised manuscript clarifies the main take-home messages of your manuscript, thanks to the correction of the simulation code, the comparison of some limit results with theoretical expectations, the improvement of several figures, and a more in-depth analysis and interpretation of your main results.

That said, Dr. Boitard and Dr. Koskela have made some additional comments and suggestions, asked for clarifications, and raised some points on your revised manuscript, that I recommend you address.

All reviewers and I consider that your manuscript addresses an important topic, with a solid and thorough approach, and I am willing to consider a revised version of your manuscript for recommendation in PCI Evol Biol. However, before making a final decision, I would also like you to consider the following concern.

First, I would like to thank you for clarifying the model description, and for providing a more complete mathematical formalism (p. 4). However, with these clarifications, I now realize that I overlooked an important modeling choice: you state that the generation (or age) G of each parent of a seed is randomly sampled from a multinomial draw, parameterized with a probability vector $\textbf{Y}^{norm}$ that depends upon the rate of dormancy. Since the draws are independent (see Lines 142-143), this amounts to considering that the two parents of a seed may not belong to the same generation (i.e., that they do not have the same age) or, to put it another way, that the ovule and the pollen of a seed were not produced in the same generation (which, biologically speaking, seems a bit odd, at least for diploid plants). If my understanding is incorrect, I recommend that you revise the description of the model, to match what the simulation program actually does; if my understanding is correct, I urge you to justify that this (biologically odd) modeling choice provides accurate results for a weak seed bank model.

Reply: The model description has been slightly modified to clarify this point. We do in fact build a pseudo-diploid model (possibly more suitable for the study of fungi than

diploid plants) which is a direct forward model version of the coalescent model by Kaj et al. (2001). We improved the description of the model and choice of assumptions on lines 124 and 146 (see below). We likewise modified the manuscript to no longer refer to the active population as "plants" or "above ground", but rather follow the terminology used by Blath et al. for other seed bank models and use "active" and "dormant" populations.

We have now added a word of caution in the discussion: "For a strict application of our model to diploid plants, future work would need to consider the constraint of having only N individual diploid parents to choose from. We expect this to likely yield slightly shorter coalescent times than in our pseudo-diploid model (based on the haploid \citep{kaj_coalescent_2001}), while our insights should still be valid."

Modifications made for the pseudo-diploid model:

Page 4 - Model description. We better separate the WF from the seed bank part of the model assumptions.

L 124 onwards describe the classic WF model:

The model represents a single, panmictic population of N hermaphroditic pseudo-diploid adults, which will henceforth be referred to as diploids for brevity. Population size is fixed and generations are discrete, so that in the absence of dormancy and selection, the population follows a classic Wright-Fisher model. In this case, at the beginning of each generation, a new individual is produced by sampling two parents from the previous generation. Once sampled, each parent contributes a (recombined) gamete to generate the new individual. Each parent is sampled with probability $\frac{1}{N}$ (multinomial sampling), leading to two vectors $\mathbf{X}_{parent 1}$ and $\mathbf{X}_{parent 2}$, containing the indices of the respective parents:

L 146 onwards describes now the seed bank model:

Once the age of each of the $2N$ parents has been determined, random individuals from the corresponding age groups are sampled (the same individual can be sampled more than once) and one recombined gamete from each of these $2N$ individuals is generated. These gametes are then randomly combined to form N new diploid individuals which constitute the current active population. Thus, the forward simulation process models two haploid dormant individuals (with different ages) which become active at the current generation and join to form a diploid individual (Figure \ref{fig_sampling}). This pseudo-diploid model formulation is implicitly equivalent to haploid gametes being resuscitated from the dormant state and fusing to create a diploid individual capable of reproduction. The probability of coalescence (p_{coal}) is therefore expected to follow haploid expectations $p_{coal}=(\frac{1}{2N}) \times b^2$). The number of recombination events is sampled from a Poisson distribution with parameter r (for example 1×10^{-8} per bp per generation).

L 22: Removed brackets (above-ground for plants) and (seeds for plants)

L 38: "We focus here on *a pseudo-diploid version of* the weak seed bank model in order to provide novel insights into the population genomic analysis of plant, fungi and invertebrate species which undergo sexual reproduction." Added the italicized part.

L 39: Removed "plant, fungi and invertebrates"

L 57: "above-ground" to "active"

L 69: "above-ground" to "active"

L 71: Removed "(above-ground for plants)"

L 73: Removed "(in pollen and ovules)" and changed "seeds" to dormant population and "above ground" to "active population"

Line 71-74: Replacing mu_s with mu_d.

L 77: Change seeds to dormant population and above-ground to active population

L 147: "above-ground" to "active"

L 283 Changed "leaving seeds dormant" to "maintaining the dormant population for longer"

L 285: "above ground plant" to "active individual"

L 291: "above-ground plants" to "active individuals"

L 313: "above ground" to "active"

L 415: "above ground" to "active"

L419: We remove the "likely realistic for plants and invertebrates" to shorten the sentence (as studied in Figure 4a in \citep{shoemaker_evolution_2018}, and \citep{koopmann_fisherwright_2017})

L 460: Removed "(in the seeds)"

L 466: Removed "plant species" for "species with dormant phase" and in same sentence "age of seed" to "age of inactive population"

L 467: We believe such studies should be done in plants, fungi and invertebrates.

L 506: removed "such as plants or invertebrates"

Reply: We refrain from referring only to plants throughout the manuscript, so as to avoid giving the wrong impression that our model is suited for only diploid plants. We try to follow the consensus in the current literature, such as the papers by Kaj et al. (2001) and the models by Blath et al. on dormancy. These authors (as well as ourselves in previous papers) often consider plants or fungi or bacteria but still refer to the dormant stage as seeds, and the dormancy model as a seed bank model.

Consistent with this concern, I suggest that you review Figure 1, as this Figure illustrates the forward-in-time two-step process considered in Kaj et al. (2001) who assumed a haploid model, and not a diploid model as in this manuscript. A revised Figure could (should) represent diploid seeds, each seed choosing its two parents from the same generation in the past.

Reply: The figure 1 has been modified to demonstrate the process of choosing two parents forward in time which produce at the current generation one diploid offspring.

Last, I invite you to consider the following minor comments/suggestions:

Line 5 of the abstract: “signals of selection” -> “signatures of selection” or “footprints of selection”

Reply: We changed it to “signatures of selection”.

Line 5 of the abstract: “more narrow” -> “narrower”

Reply: We changed it to “narrower”.

Keywords: “weak, dormancy” -> “weak dormancy” (remove the comma)

Reply: The comma is removed.

Lines 15-16: “spatial structure” -> “genetic structure” or “genetic differentiation”

Reply: We changed it to “genetic structure”.

Line 23: “the common ancestor of a population” -> “the common ancestor of a sample of genes from the active population”

Reply: We changed it according to the suggestion.

Line 49: “representing the populations of size N from the past” -> “representing the past populations of size N”

Reply: We changed it according to the suggestion.

Lines 79-80: “only the non-dormant lineage is affected by recombination” -> “only the non-dormant lineages are affected by recombination”

Reply: It is now plural.

Line 127: “indicies” -> “indices”

Reply: The typo is corrected.

Line 138: Completing Dr. Jore Koskela's first comment (see below), the definition of the probability vector $\textbf{Y}^{norm}$ should read: $\textbf{Y}^{norm} = \frac{\textbf{Y}}{\sum_{j=1}^{m} Y_j}$, or: $Y_k^{norm} = \frac{Y_k}{\sum_{j=1}^{m} Y_j}$

Reply: Thank you, we chose the first suggested option.

Line 154 "Sequantially" -> "sequentially"

Reply: Thanks for spotting this typo.

Line 187: "As previously stated the simulation process can ran, independently of tskit, but is required when planning to analyze the genealogy" -> "As previously stated the simulation process can be ran independently of tskit, but the latter is required when planning to analyze the genealogy."

Reply: The sentence has been corrected.

Lines 191-192: "(Figure S1 and Table S1 for empirically sufficient number of calibration generations given for a recombination rate)" -> "(see Figure S1 and Table S1 for characterizing empirically the sufficient number of calibration generations needed for a given recombination rate)"

Reply: The part in the brackets has been modified to account for your suggestion.

Line 221: "if it stays at a size of 2N for m consecutive generations" -> "if its number of copies is 2N for m consecutive generations"

Reply: Thank you, changes were made.

Lines 285-286: "[...] decreasing the value of b (i.e. the longer seeds remain dormant)" -> "[...] decreasing the value of b (i.e., leaving seeds dormant longer)"

Reply: The bracket part has been modified. But we used the british way of leaving out the comma after i.e.

Figure 2b: indicate what shaded areas represent (as in Figures 3b and 3c).

Reply: A description has been added: "Shaded areas represent the 95 \% confidence interval."

Figure 3: The last sentence in the legend ("Dashed-blue lines indicate theoretical expectations of a Ne-scaled population corresponding to a given seed bank strength.") repeats what is written above ("The blue lines indicate the time to fixation in a population without dormancy but with an effective population size scaled by"), does it not?

Reply: We removed the duplicated sentence.

Line 331: "(e.g. Figure 4b, 4d and S10)" -> "(e.g., Figures 4b, 4d and S10)"

Reply: We added the “s”.

Figure 4: The legend for (3c) should rather read: “(c) π assuming two recombination rates r = 10−7 per bp per generation (c1) and r = 5 × 10−8 per bp per generation (c2).”

Reply: This part of the legend has been adapted according to your correction.

Figure 4: “(b) Normalized π as divided by the average neutral branch diversity from (a) using the values 2,000 and 16,000 for b = 1 and b = 0.35, respectively.” What are the values 2,000 and 16,000? (values of what?) It does not seem to relate to a scaling by $\frac{1}{b^2}$, and I am therefore not quite sure to understand how π is normalized.

Reply: It now reads as follows: “(b) Normalized π as divided by the average neutral branch diversity, namely approx. 2000 for $b=1$ and approx. 16000 for $b=0.35$ (see (a) or (c) between sequence range of 0 to 0.2×10^6 or from 0.8×10^6 to 1×10^6).” Further, 1/(0.35*0.35) is approx. a scaling factor of 8, which is how the scaling of ~16000 can also be calculated.

Lines 327-328: “Moreover, stronger dormancy also generates narrower selective sweeps around sites under positive selection which have reached fixation”: when comparing Figures S10 and 4, I would rather conclude that stronger selection (i.e., s = 1 vs. s = 0.2) generates narrower selective sweeps around sites under positive selection (since the rate of dormancy is the same in both Figures). Please, correct the sentence, or be more explicit.

Reply: The sentence has been corrected to “Moreover, comparing the width of the selective sweeps valley of polymorphism in presence and absence of dormancy, we conclude that stronger dormancy generates narrower selective sweeps around sites under positive selection which have reached fixation (Figures \ref{fig:sweep1}, \ref{fig:sweep3} and \ref{fig:largeseg}b)”

Figure 5: in the legend, (a12,b12) should read (a21,b21) to be consistent with the Figure.

Reply: It has been corrected now.

Line 370-372: I would move the sentence “Following the classic procedure to detect sweep [...]” above, since it provides the criteria used to obtain the results from lines 367-370.

Reply: Thank you, we moved the sentence according to your suggestion.

Lines 379-380: “We note that there is a much sharper decrease in the rate of detection of false positive sweeps (neutral simulation line in Figure 5) under seed bank compared to the absence of a seed bank.” Do you have any insight about why would that be? It might seem a bit counterintuitive.

Reply: We added “[…] , likely being a direct consequence of the increased linkage decay around the site.”

I thank you very much for submitting your manuscript to PCI Evol Biol, and I look forward to receiving your revised manuscript.

Best regards,
Renaud Vitalis

Reviews

Reviewed by Jere Koskela, 27 Feb 2023 10:08

I think this is a thorough paper on a topic which is of both mathematical and biological interest, and the comments I made in the previous round have been fully addressed. I only have two further, superficial comments to add:

1. Page 4, line 137: In the vector Y = (Y_1, ..., Y_m), it looks as if Y_k means the kth entry of Y. However, on the same line Pr(Y_k) seems to stand for the probability of the event that a random variable takes the value k. Do you mean Pr(Y_j = k)?

Reply: Indeed, the line has been changed according to the suggestion of Renaud Vitalis (see above).

2. Page 6, line 187: "As previously stated the simulation process can ran..." seems to be missing a "be".

Reply: The “be” has been added.

Reviewed by Simon Boitard, 08 Mar 2023 21:22

The authors have made substantial efforts to tackle the questions raised by all reviewers, including myself. They corrected some errors in the code and updated the simulation results, they included new analyzes on allele frequency dynamics and sweep detection, and they significantly revised the text and figures of the manuscript. Some clarifications and corrections are still needed but should be very easy to implement.

Line 27, 53 and many others : ‘coalescent’ is the name for the stochastic process describing the genealogy of a sample, so it is correct to talk about the Kingman ‘coalescent’ (for instance). But the events within this process are ‘coalescence’ events, and similary one should talk about ‘coalescence’ rates, ‘coalescence’ times …

Reply: We corrected it to “coalescence” at the respective lines.

L48-50 : Is it unclear what ‘representing the populations of size N from the past’ means.

Reply: We rewrote it according to one the reviewers suggestion to: “representing the past populations of size N”

L85 : ‘a’ Sequentially Markovian Coalescent.

Reply: Fixed.

L89-94 : I don’t understand the arguments here. If the fixation time of selected alleles is not multiplied by $1/b^2$ but by a smaller rate, it is smaller than expected and the selection is thus more efficient, not ‘altered’. Similarly at line 97, why do seed banks ‘decrease’ the efficacy of

selection. Based on the results of Figure 3, I assume that L91 is maybe wrong and that fixation time is multiplied by a ‘higher’ rate.

Reply: the last part of the sentence on line 91 was ambiguous. We now rephrase it as “seed bank model with positive selection and show that the time to fixation is not multiplied by $1/b^2$ (as for neutral alleles) but by higher factor (between $1/b^2$ and $1/b$)”

L100 : ‘on’ is probably missing between ‘seed bank’ and theta’.

Reply: We added the “on”.

L137 : Yk is a probability, so one should write Y_k=b(1-b)^(k-1), not P(Y_k)=…

Reply: We changed it according to your suggestion.

L154-156 : I don’t understand where SMC is used, given that simulations are forward in time and SMC is a backward model.

Reply: We removed the sentence “In other words, we use the sequentially Markovian Coalescent approximation of the Ancestral Recombination Graph, \citep{mcvean_2005_SMC}).” Indeed, we don’t use the SMC anywhere, we just mention a SMC method that has been used for inference of the weak dormancy germination rate parameter, where the same assumptions about the occurrence of mutation in the seed bank has been made, as justification for modeling mutations the same way.

L188 : Is it meant that ‘tskit is required when planning …’ ?

Reply: When planning to do anything genealogy-related then it is necessary (also for calculating branch-based statistics). When only being interested in fixation times or probabilities, then it is not required for planning.

L190 : Since this calibration phase strongly dpends on the population size, it would be better to describe it after introducing the simulated population sizes.

Reply: We merged the paragraphs and moved the explanation of the calibration phase after introduction the population size as suggested.

Figure 2, S1 and S8b : Why recording the ‘oldest TMRCA per sequence’ ? This switches the focus from the mean coalescence time to the maximum coalescence time, which probably explains why this quantity depends on the recombination rate (Figure S1).

Reply: We agree and getting an intuition about the maximum coalescence time further helped to choose a suitable *calibration phase*. Working with fully coalesced genologies was required for calculating branch based statistics for studying the

sweep signatures. For the main figure normalization we removed the effect of recombination, which is in line with your intuition.

Line 307 and 311 : I am not sure that 3c is the correct reference here, is it not 3b ?

Reply: Yes, we corrected it.

Line 317 : Is it not rather ‘decreasing b’ ?

Reply: Indeed, yes. It has been fixed.

Line 328 : ‘S10’ → ‘(Figure S10)’

Reply: Thank you for spotting this.

Line 345 : ‘smooth’ would maybe fit better than ‘sharp’ ?

Reply: Indeed, it should be smooth.

Line 365 : ‘these variations’

Reply: We fixed it.

L372-375 : The link between the two sentences is not obvious to me. Decreasing the window size may increase the rate of true positives more than that of false positives, resulting in a higher power for a fixed false positive rate threshold.

Reply: That is what we meant by decreasing sensitivity. We modified it to: “Decreasing the window size is generally associated with a loss of sensitivity, increasing the rate of true and false positives.”

The old version: “Decreasing the window size is generally associated with a loss of sensitivity, increasing the rate of false positives.”

Figure 5 : ‘0 generations’ is missing at line 4 of the legend.

Reply: It has been added.

L375-376 : It seems to me that these sweeps are already detectable in b21, not only in b22.

Reply: We modified the sentence to: “However, the detectability of older sweeps ($>$2,000 generations) is increased for $b=0.35$ (Figure \ref{fig:detect_sweeps} b22).”

L414-417 : A verb is missing in this sentence.

Reply: The sentence has been corrected.

L 425-428 : This sentence is a bit confusing, especially the part ‘which did not compute the probability of fixation of an advantageous allele’.

Reply: Indeed, the sentence was wrong. It now reads: “Our results thus mitigate the previous claim that (weak) seed banks may amplify selection, making it relatively more efficient with regards to the effects of genetic drift, *while it does not alter the probability of fixation of an advantageous allele.*”

Figure S3 : The legend says that time is normalized, but it is not clear which normalization is used.

Reply: Legend has been modified by: “[...] time normalized by the $b=1$ estimate for each respective selection coefficient”

Figure S8 : « =0.35 » is missing in the legend (current text ‘b=1 and b for ... ».

Reply: It has been added.

Figure S9 : I don’t understand why Ne*s is 800 for b=1 but 400 for b=0.35. I thought the aim of this analysis was to have the same Ne for both values of b ?

Reply: I think, the 800 is a typo and should be 400:

b=1: 2000 * 0.2 = 400; b=0.35: 245 * 1/(0.35*0.35) * 2 = 400

We changed it in the legend.

Figure S10 : Same comment as for S9, although I don’t clearly see whether these two figures have the same or different objectives. I note that in Figure S9 the sweep is narrower with b=1, while in S10 it is narrower with b=0.35 : how can we explain this ?

Reply: Figure S9 and S10 have different objectives. In Figure S9 we correct for Ne by choosing N=245 with b=0.35 and N=2000 for b =1. It is the only Figure in the paper for which we change the N when comparing seed bank and non-seed bank simulations, otherwise it is N=500 throughout the paper. In S10 our objective was to have the same N=500 as in the other Figures of the paper, but we increased the sequence length to 10MB. And in Figure S10 we choose a strong selection coefficient, leading to this rather narrow signature.

Reviewed by William Shoemaker, 15 Mar 2023 07:21

The authors have done an admirable job improving the manuscript, both by addressing my comments and the comments of my fellow reviewers. I have no further comments and believe that it should be recommended by PCI Evol Biol.

Reply: We appreciate that, thank you.