Calculating hazard rates of introgression with branching processes

Ghosh, A.

CHAPTER 4: QUANTIFYING INTROGRESSION RISK WITH REALISTIC POPULATION GENETICS

To be resubmitted

Abstract

Introgression is the permanent incorporation of genes from the genome of one population into another. This can have severe consequences, such as extinction of endemic species, or the spread of transgenes. Quantification of the risk of introgression is an important component of GM crop regulation. Current introgression models disregard important factors such as genetical mechanisms, repeated invasions, and stochasticity. We present a method to quantify introgression risk that incorporates all these crucial aspects. This is done by combining two modelling approaches that are traditionally separated: population genetics, and branching process theory. We calculate a probabilistic risk measure for introgression, called the hazard rate. When the recipient population is small, drift dominates, and simulations of population genetic models are required to calculate the hazard rate, whereas in large populations selection drives introgression, and efficient numerical procedures based on branching process models suffice. We illustrate this by studying the effects of linkage and recombination on introgression risk at different population sizes.

1. Introduction

Human activity has dramatically increased the rate of hybridisation between species or ecotypes by agriculture, trade, and travelling. An important consequence is the potential occurrence of introgression, when genes from the genome of one population or species become permanently incorporated into the genome of another. Hybridisation and introgression can have undesirable effects, such as the extinction of endemic species, or the spread of resistance genes, which may for instance result in an increased weediness of plant species [1,2]. Especially the application of genetically modified crops in agriculture has raised many concerns about the incidence of introgression. Quantification of the risk of transgene introgression is therefore a key component of the regulation of GM crops.

Introgression processes typically have two major characteristics: hybridisation occurs recurrently, and at least initially the fate of invaders is highly capricious, due to chance events. Both aspects are generally ignored in introgression models. Many studies concern deterministic models with single invasion attempts [3]. The relative fitness advantage of an invading gene is then used to measure invasion risk. This measure is closely related to the probability of success of a single invasion in an infinitely large population [4,5], or the fixation probability of single mutations in finite populations [6].

There are several reasons why the probability of fixation is an inappropriate measure of introgression risk. First, when the repetition of invasions persists indefinitely, the foreign gene will eventually become fixed in any population of finite size due to genetic drift, regardless of its fitness effects. Models based on single invasions, however, predict that the establishment probability of deleterious invading genes is zero, for infinite populations, or very small, when population size is finite. Similarly, repetitive invasions of advantageous genes in (infinitely) large populations will have a success probability of one, even if the success probability of a single invasion is very small. Second, even when repetitive invasions only occur during a finite time period, the establishment probability of the invading gene is much higher than would be predicted on the basis of the single invasion scenario. Third, because in most cases the initial number of invaders is small, invasion attempts usually fail several times due to demographic stochasticity, before permanent establishment is initiated. The time until this initiation is an important characteristic of invasion risk, that should be included in its quantification.

A proper measure of introgression risk should be based on the probability that a successful invasion (the initiation of a permanent introgressed lineage) occurs within a given period of time. We previously developed methods to calculate such probabilities, based on stochastic population dynamic models [7,8], where we assumed an (infinitely) large receiving population. In the current paper we generalize our methods to include invasions in small to medium-sized populations. Another factor of critical importance for introgression risk is the location of an invading gene in the crop genome. Linkage to a crop gene that is under positive selection in the wild population may considerably enhance introgression risk, whereas linkage to a deleterious gene will reduce it. Multi-locus genetics constitute another important aspect of introgression risk that has been ignored in the modelling literature until now [9]. Specifically the use of genomic linkage as a strategy to mitigate introgression is still in the conceptual stages [10]. Traditional population genetic models of selection and recombination are inappropriate for these purposes since they only consider evolutionary time scales. In these models new genotypes are created through mutation and the time between successive mutations is assumed to be very large compared to the generation time. Therefore each new mutation can be considered as a single invasion into a stable population, and the effects of repeated invasions are ignored. In models of hybridisation, such as we consider here, invasion repetition becomes an important element that changes the population genetic dynamics considerably.

We present here a method to incorporate linkage, recombination, and invasion repetition into population genetic models, and quantify their effects. The methods that we developed previously [7] cover situations with simple single locus two allele dynamics. In this paper, we show how these can be generalized to a multi-locus system, exemplified by a two-locus two allele situation. We use this model to study the effects of linkage between a fitness enhancing (trans)gene and a domestication gene with deleterious effects under natural conditions.

We have previously shown how the hazard rate of introgression can be used as a measure of introgression risk [7]. This hazard rate is defined as the probability per unit time that a so-called 'introgression event' occurs, given that it has not previously occurred. In our previous models, we considered (infinitely) large wild

populations, where the probability of interaction between hybrid lineages can be ignored. In that case, an introgression event corresponds to the initiation of a permanently introgressed lineage. In finite populations this definition is problematic, however, since individuals from different lineages may produce offspring together. Therefore, permanent introgression is not necessarily initiated by a single lineage. One possibility would be to study the frequency of the transgene after a certain period, as done in [11]. The time at which this frequency exceeds a given level could then be considered as the starting point of introgression. However, since the choice of the threshold frequency is arbitrary, this is also a doubtful definition. Here, we propose an alternative definition: permanent introgression has been initiated at or before a specific time if the invading allele will go to fixation in the population even if no further invasions occur after that time. This definition is equivalent to the one we used before for the situations that we studied previously. For more complicated cases, such as considered presently, these probabilities can be calculated from simulations by means of survival analysis methods e.g. [12]. This implies that, with the methodology presented in this paper, hazard rates can be calculated from simulations of stochastic population dynamic models with any degree of genetic and/or ecological complexity, including models for small or medium sized populations.

The hazard rate can also be calculated from branching process models, using the approach that we developed previously for a single-locus system [7]. Whereas this is an approximation, based on infinitely large population sizes, it has the advantage that it does not require computer simulations, but only the numerical solution of a system of equations, which is a much more efficient method. As an illustration, we show here how to apply this method to situations with two-locus two-allele dynamics. The generalization to more complex cases is straightforward. For the model we consider here, the branching-process approximation already works extremely well for population sizes of about 100 individuals.

We discuss our methods in the context of transgene introgression from GM crops into wild populations, but they can be used in any situation where repeated invasions occur. For example, they may have important applications in the evolutionary dynamics of microbial systems, where mutation rates are high, and additional modes of genome modification occur, such as bacterial competence [13]. Other examples are epidemic processes [14], exotic species invasions [15] and the origin, growth, and spread of tumours [16].

2. The model system

We consider situations with a flow of pollen from a crop field into a nearby population of a wild relative. The crop contains a transgene conferring a positive fitness effect, which is physically linked to a domestication gene with a negative fitness effect. In heterozygotes, recombination can cause the transgene to become uncoupled from the domestication gene, creating the haplotype with the highest fitness. The crop is assumed to be homozygous at the transgene and the domestication gene loci; the wild population is assumed to be homozygous for the wild-type alleles at both loci, and these alleles are taken to be selectively neutral. We represent the transgene and domestication gene alleles by the capital letters 'A' and 'B' respectively. The corresponding wild-type alleles are given by the lower-case

letters 'a' and 'b' respectively. All plants in the model are hermaphroditic annuals, and mate randomly. The wild population is assumed to have fixed size.

2.1. **Branching process approach.** This approach is analogous to that described in [7]. The wild population is assumed to be large enough such that hybrids and their descendants initially only cross with wild individuals. Consequently, there are no hybrid-hybrid crosses, which means that homozygotes for either the transgene or domestication allele will not appear in the invasion analysis using the branching process approach. While it is true that such homozygotes are eventually produced in reality, we only concern ourselves with the initial phase of the invasion when the numbers of invaders is still small. This assumption allows hybrid lineages to be considered independent of each other, which simplifies the dynamics considerably.

We assume that a Poisson distributed number (m) of hybrids (genotype $ABab$) is produced per generation. These hybrids produce a number of offspring according to a Poisson distribution with mean $2w_{ABab}$. The genotypes of these offspring depend on the recombination rate (r) between the transgene and the domestication gene. With probability $\frac{1}{2}r$ the genotype is $Abab$, with probability $\frac{1}{2}r$ it is $aBab$, with probability $\frac{1}{2}(1-r)$ ABab, and with probability $\frac{1}{2}(1-r)$ it is $abab$. As mentioned, we assume that the population is large, so that these four genotypes are the only ones we have to take into account. Furthermore, we only have to consider the fate of the two genotypes that carry the transgene (i.e. $ABab$ and $Abab$), since the others cannot initiate lineages that lead to the permanent introgression of the transgene.

We use the symbol wi to denote the fitness of an individual of genotype-i, with wild genotypes ($abab$) having a fitness of 1. Note that a fitness of one corresponds to an individual producing on average two offspring, since each parent only contributes half of an offsprings chromosomes. For instance, an individual of genotype $Abab$ produces a Poisson-distributed number of offspring with mean $2w_{Abab}$. Since mating only occurs with a wild type ($abab$), these offspring have genotype $abab$ or $Abab$ with probability 0.5.

2.2. **Population genetic simulation approach.** The simulation-based approach considers a wild population of a fixed size, N, which may be small. Consequently, hybrids may mate with other hybrids, and all possible genotypes may appear. The population is therefore represented by a vector of length 16, where each component corresponds to the number of individuals of a given genotype. The vector has length 16 rather than 9, because we distinguish chromosomes inherited from the mother from those of the father. The life-cycle progresses according to three stages: reproduction, death of adults, and germination of seeds. Reproduction takes place through the production of exactly N seeds. For this, first the expected frequency of the 16 genotypes in the following generation is calculated given random mating, their current frequency, and the fitness effects of the two loci. Then the frequencies of the 16 genotypes among the produced seeds are drawn as random numbers from a multinomial distribution with the expected frequencies as probabilities. A small number of randomly selected seeds is then replaced by seeds created through hybridisation between the wild population and the crop. For this, the paternal contribution of the selected seeds is replaced with an AB-gamete from the crop.

The number of hybrids produced is chosen by drawing a random number from a binomial distribution with N trials and a m/N success probability. In the limit of large N, this becomes a Poisson distribution with a mean of m, which agrees with the use of a Poisson distribution in the branching process approach. After these seeds have been created, the adult individuals die, and all seeds germinate and establish themselves. The simulation model was programmed and run in the programming package R. For every combination of parameters settings, we ran 100,000 replicates for each value of n between 1 and 100.

2.3. **The hazard rate.** An introgression event has occurred at or before a specific time T if the transgene will go to fixation even when no further hybridisation occurs after T. The hazard rate of introgression is defined as the probability that an introgression event occurs at a time n given that it has not occurred before. The use of hazard rates is well established in the field of survival analysis, where they represent instantaneous mortality risks. The interested reader can find more information on this in [12] for example.

The hazard rate can be expressed as follows:

$$H(n) = P(T = n \,|T > n-1) = \frac{P(T = n)}{P(T > n-1)} \tag{1}$$

It is therefore a function of time that can be calculated from the distribution of T.

2.4. **Calculating the hazard rate from branching processes.** When invasion occurs continuously, the hazard rate reaches a positive asymptote (as depicted in Fig. 1). It is this asymptote that we use as a measure of introgression risk. Details of the derivation are given in the Appendix. Here we summarize the main results. According to branching process theory (see the Appendix and e.g. [17]) the extinction probability of a lineage after a single invasion of the genotype *Abab* equals the smallest root of the following equation:

$$q = e^{-w_{Abab}(1-q)} \tag{2}$$

where w_{ABab} represents the fitness of an individual of genotype-ABab. The smallest root q will be less than one when w_{Abab} is greater than 1. The asymptotic hazard rate is given by:

$$\hat{H}(q) = 1 - e^{-m\left(1-\hat{f}_{I_{ABab}}(q)\right)} \tag{3}$$

where $\hat{f}_{I_{ABab}}(q)$ satisfies the following equation:

$$\hat{f}_{I_{ABab}}(q) = e^{-w_{ABab}\left(1-(1-r)\hat{f}_{I_{ABab}}(q)-rq\right)}. \tag{4}$$

with r the recombination rate between the loci of the transgene and domestication gene. Equations (2) and (4) can be solved numerically.

2.5. **Calculating the hazard rate from population genetic simulations.** To estimate the hazard rate using population genetic simulations, we need to find the probability that an introgression event has occurred at each time step. This is done by running the simulation model with continuous hybridization for n generations. At that time hybridization stops and the simulation continues until the transgene is either fixed or has disappeared from the population. The proportion of replicates

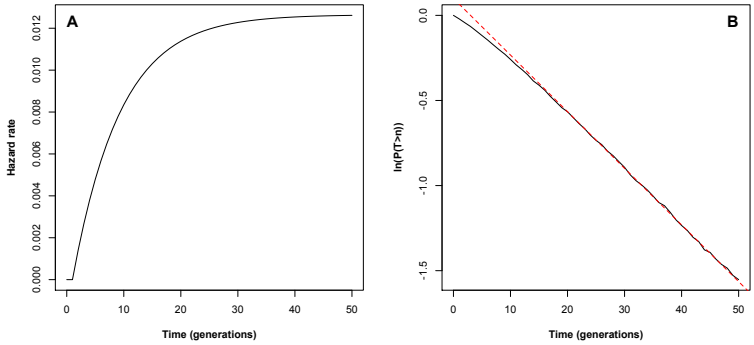

FIGURE 1. Calculation of the hazard rate for both models. A) The hazard rate plotted against time, as calculated by the branching process method. The asymptote reached corresponds to the value given in (3). B) A sample output from the simulation based method. $\ln P(T > n)$ plotted against time (black line). The slope of a regression fitted on the linear part of the curve (red line) is used to estimate the asymptotic hazard rate. Here the linear part is taken to start after 15 generations. For both plots, the curves are plotted as lines for the sake of clarity, even though the model system is in discrete time. Used parameters are$w_{ABab} = 0.9$, $w_{Abab} = 1.2$, $r = 0.005$, $m = 1$. For the simulations, $N = 10$.

that reach fixation after n generations of hybridization provides an estimate of the probability . From these probabilities, the hazard rate can be calculated at each value of n, using Eq. (1). For large values of n, the probabilities in Eq. (1) can be very low, giving a large error in the estimation of the asymptotic hazard rate. However, a more accurate estimate of the asymptotic hazard rate can be obtained by taking the complement of the exponent of the slope of the linear regression of $\ln P(T > n)$. Since the asymptotic value of the hazard rate is not reached straight away, the location where the linear part of this function starts has to be determined. This can be done by eye, or through formal methods for estimating lag-times in exponential distributions (see e.g. [18]). Note that this has to be done separately for every combination of parameter settings, since the rate at which the asymptote is approached may differ between settings (see Fig. 1).

3. RESULTS

Figure 2 shows the effect of fitnesses of type-$ABab$ and $Abab$ individuals on the asymptotic value of the hazard rate. As expected, the asymptotic hazard rate increases with increasing fitness for both genotypes. The slope of the increase depends both on the population size and the recombination rate. Figure 2 also shows that there is a strong interactive effect of the recombination rate and the population size. At a low recombination rate ($r = 0.005$) the hazard rate is generally the highest for the smallest population size ($N = 10$). On the other

hand, at higher recombination rates ($r = 0.05$, $r = 0.5$), the same population size generally gives the lowest asymptotic values for the hazard rate. The effect of changing the fitness of type-*Abab* individuals is larger at high recombination rates than at lower recombination rates.

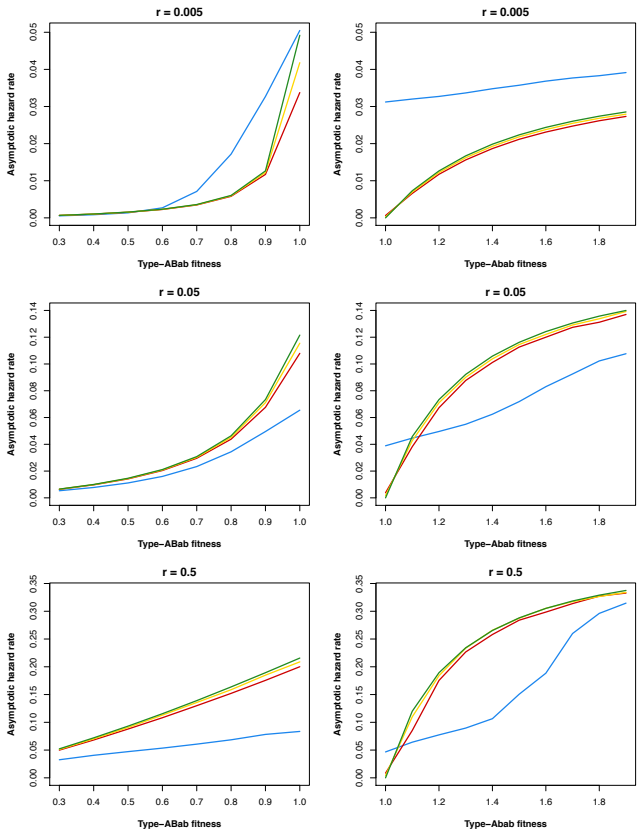

FIGURE 2. The asymptotic hazard rate as a function of the fitness of type-*ABab* individuals (left), and against the fitness of type-*Abab* individuals (right hand), for three different recombination rates. For the left column, m_{ABab} is varied from 0.3 to 1.0, with w_{Abab} set to 1.2; for the right column, w_{Abab} is varied from 1.0 to 1.9, with w_{ABab} set to 0.9. For all plots, $m = 1$. Results for the branching process are shown in green. Simulation results are shown for three different population sizes: 10 (blue), 50 (red) and 100 (yellow).

The interactive effects of recombination rate and population size is studied in more detail in Fig 3. For the branching process, an increase in the recombination rate simply results in an increase in the asymptotic hazard rate. In the simulation model, the hazard rate reaches very high levels in extremely small populations, consisting of just a few individuals. At intermediate population sizes, of about

10-20 individuals, however, the hazard rate is quite low. At large populations, the hazard rates of the simulation-based method approach the branching process value. As can be seen from the figure, the branching process approximation already works well at population sizes of 100 individuals, especially with low recombination rates.

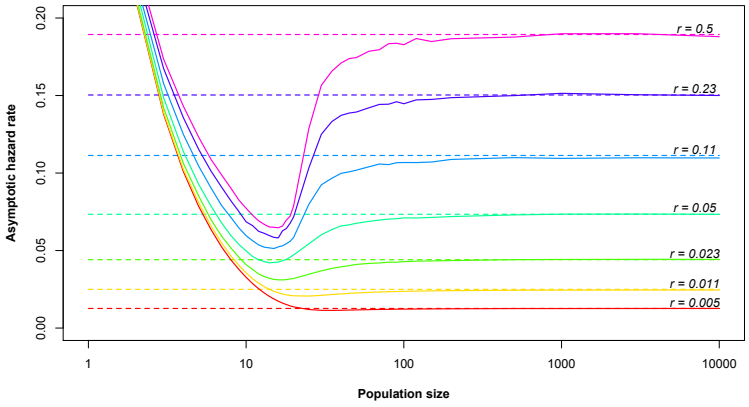

FIGURE 3. The asymptotic hazard rate plotted against population size (on a logarithmic scale) for $w_{ABab} = 0.9$, $w_{Abab} = 1.2$, $m = 1$, and various recombination rates in different colours as indicated in the plot. The branching process results, which are independent of the population size, are shown as dotted lines, the simulation results are shown as solid lines.

4. Discussion

The hazard rate provides an intuitive and accurate way to quantify introgression risk with repeated invasions. In previous papers we demonstrated how this measure can be used to study introgression risk in relation to fitness parameters [7] and crop management schemes [8]. The underlying genetics, however, did not go beyond one-locus two-allele situations. With the methods presented in the current paper, more realistic population genetics can be incorporated into hazard rate calculations. For sake of simplicity we demonstrated this by a two-locus two-allele model, but generalization of our methods to account for more complicated genetic mechanisms is straightforward. Similarly, more complex life cycles or ecological conditions can be incorporated. Thus, the current generalization makes it possible to calculate hazard rates for realistic models with a high level of complexity.

We presented two complementary approaches for hazard rate calculations. For small populations, where drift plays an important role, simulation models should be used to calculate hazard rates. Branching process models provide a fast and efficient way of calculating hazard rates for large populations, where simulations require much time and computer space. Since the branching process approach assumes infinitely large population sizes, and neglects the possibility of interactions between hybrids, the calculated hazard rate based on this approach is an

approximation for situations with finite population sizes. As demonstrated by our results, however, this approximation is very efficient, and may already work well at moderate population sizes, of 100 individuals or more.

As seen in Fig. 3, the effects of drift can be counter-intuitive and cannot be extrapolated from the results of branching-process models. While one might expect that drift always increases the probability of a successful invasion, in some intermediate populations, invasion risks are smaller than that at larger populations. The reason for this lies in the fact that selection is the main driving force behind invasions in larger populations. Under strong selective pressures, the invasion risk can be high at large populations, but smaller at small populations because potential invaders can be removed from the population by drift.

Another difference between the results of the two approaches occurs when the fitness of the introgressed *Abab* genotype is equal to one, i.e. when the transgene does not have any fitness effect. In this case, the hazard rate in the branching process model is equal to zero. This is because at neutrality, any increase in the frequency of the transgene is caused only by genetic drift, which is absent in the large population size assumed by the branching process. In the simulation model, drift does occur and therefore the asymptotic hazard rate for this model need not approach zero with these identical parameters. Consequently, the discrepancy between the models is largest at small populations.

The results from both approaches show that recombination rates between loci have large effects, and thus that linkage is an important aspect of introgression modelling. At high recombination rates, there are more type-*Abab* individuals produced, and so changing the fitness of such individuals has a larger effect than at low recombination rates (see Fig. 2). The sensitivity of the hazard rate to recombination rates is smallest in small populations, where introgression is primarily driven by drift (see Fig. 3).

We find high hazard rates at small population sizes that are of the same order of magnitude as the hybridisation rate. This is because hybridisation alone is enough to push invading genes to fixation. Consequently, many copies of the domestication gene also go to fixation under these circumstances. The hazard rate typically reaches a minimum at population sizes of the order 10-20. This is because introgression is primarily still driven by drift in these circumstances, and the number of invaders is small compared to the number of residents, so drift acts to push these invaders out of the population. At larger population sizes, of the order of 80 and higher, selection becomes the dominant factor, and the results from numerical simulations approach those predicted by branching processes.

These results shed light on the circumstances when branching process models can be used as good predictors for invasion risk, and when simulation models should be used. Many previous attempts to model invasions using branching processes had to consider an invasion into a large resident population (e.g. [19,20]). As of now, little work has been done in investigating how large a resident population is necessary for the results of such models to hold (but see [21]). Our results suggest that branching processes are valid for population sizes that are ecologically relevant.

While the approaches outlined are important for calculating introgression risks, there is still much to be done. In the simulation model, population sizes were

assumed to be fixed but it would be more biologically relevant if this assumption were relaxed. Also, while we considered introgression into a single wild population, we have not taken into account the metapopulation structure of wild populations, which would be an important aspect of a more complete model [22]. Generalising approaches from metapopulation ecology [23] would be an important step to take. Another extension is to consider time-inhomogeneous processes, such as caused by e.g. crop management schemes, as considered in [8].

In conclusion, hazard rates provide an important characterisation of invasion risk in situations with repeated invasions. They are applicable through a range of different modelling frameworks, and provide an intuitive measure of risk in a complex stochastic process.

Appendix A. Appendix

A.1. **Derivation of (2).** The lineage initiated by an individual of genotype-*Abab* becomes extinct if and only if all of its offsprings' lineages become extinct, or if it produces no offspring. Using this logic, we find the following:

$$\begin{aligned} q &= \sum_i P\left(\xi_{Abab} = i\right) \sum_{j=1}^{i} \binom{i}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{i-j} q^j \\ &= \sum_i P\left(\xi_{Abab} = i\right) \left(\frac{1}{2}\left(1+q\right)\right)^i \\ &= G_{Abab}\left(\frac{1}{2}\left(1+q\right)\right) \end{aligned} \tag{5}$$

where ξ_{Abab} represents the number of offspring produced by a single individual of genotype-*Abab*. The factors of $\frac{1}{2}$ arise because only half of the individuals offspring will be of type-*Abab*, with the other half being of type-*abab*. In the last line, we have used the definition of a probability generating function (p.g.f.), and $G_{Abab}(s)$ represents the p.g.f. of the offspring production of an individual of genotype-*Abab*. Recall that the definition of a p.g.f. of a random variable Z is defined as $E[s^Z]$, where s takes values in [0,1]. Using the assumption that the offspring of all individuals are Poisson-distributed, we can use the Poisson form of a p.g.f. to write (5) as follows:

$$q = e^{-\frac{m_{Abab}}{2}(1-q)} \tag{6}$$

where m_Abab represents the average number of offspring of an individual of genotype-*Abab* made through both male and female sexual components of the plant. The expression of a Poisson p.g.f. is often used in branching processes, and can be found in [17]. We take the average number of offspring as twice our used value of fitness, which leads to the expression that $\frac{1}{2}m_{Abab} = w_{Abab}$. Using this in (6) results in the required expression in (2).

A.2. **Derivation of (3).** To model the repeated invasion of individuals, it is helpful to introduce a so-called type-0 individual into the branching process model. Each generation, a type-0 individual produces a random number of hybrids, and exactly one of itself. This is a convenient tool for modelling immigration in branching process, see e.g. [17].

Using multi-dimensional p.g.f.s simplify the derivation of (3). We define the following multi-dimensional p.g.f. of the offspring of a single type-i individual, $i \in \{0, ABab, Abab\}$

$$F_i\left(s_0, s_{ABab}, s_{Abab}\right) = E\left[s_0^{Z_0(1)} s_{ABab}^{Z_{ABab}(1)} s_{Abab}^{Z_{Abab}(1)} \left| Z_i\left(0\right) = 1, Z_j\left(0\right) = 1 \text{ for } j \neq i \right.\right] \tag{7}$$

where $Z_i(n)$ denotes the number of type-i individuals at time n.

The definition from (7) combined with the model assumptions described in the text result in the following:

$$F_0\left(s_0, s_{ABab}, s_{Abab}\right) = s_0 G_0\left(s_{ABab}\right) \tag{8}$$

since a type-0 individual produces one of its own type, and a random number of type-$ABab$ plants according to a p.g.f. $G_0(s)$.

$$F_{ABab}\left(s_0, s_{ABab}, s_{Abab}\right) = G_{ABab}\left(\frac{1}{2}\left(r s_{Abab} + \left(1-r\right) s_{ABab} + 1\right)\right) \tag{9}$$

since a type-$ABab$ individual produces offspring according to a p.g.f. $G_{ABab}(s)$, of which a proportion $\frac{1}{2}r$ is type-$Abab$, a proportion $\frac{1}{2}(1-r)$ of type-$ABab$ and the remaining proportion of $\frac{1}{2}$ are other types.

Now we introduce the random variable $I_i(n)$, $i \in \{0, ABab, Abab\}$, which is defined as the total number of type-$Abab$ individuals produced with a type-$ABab$ parent, in the lineage initiated by a single individual of type-i. We can manipulate the p.g.f. of $I_i(n)$ to coincide with the joint p.g.f. shown in (7) as follows:

$$\begin{aligned} f_{I_i(n)}\left(s\right) &= E\left[s^{I_i(n)}\right] \\ &= E\left[E\left[s^{I_i(n)} \left| Z_0\left(1\right), Z_{ABab}\left(1\right), Z_{Abab}\left(1\right)\right.\right]\right] \\ &= E\left[E\left[s^{\sum_{k=1}^{Z_0(1)} I_0(n-1)^{(k)} + \sum_{k=1}^{Z_{ABab}(1)} I_{ABab}(n-1)^{(k)} + Z_{Abab}(1)}\right.\right. \\ &\qquad \left.\left. \left| Z_0\left(1\right), Z_{ABab}\left(1\right), Z_{Abab}\left(1\right)\right.\right] \left| Z_i\left(0\right) = 1, Z_j\left(0\right) = 0 \text{ for } j \neq i\right.\right] \end{aligned} \tag{10}$$

where the random variables $I_j(n-1)^{(k)}$ represent the total number of type-$Abab$ individuals produced up to and including the next $n-1$ generations by type-$ABab$ individuals in the lineage initiated by the kth individual of type-j from the first generation. We can use the fact that individuals in the branching process reproduce independently and that individuals of the same type have identical offspring distributions to rewrite the right-hand side of (10) as follows:

$$\begin{aligned} &f_{I_i(n)}\left(s\right) \\ &= E\left[E\left[s^{I_0(n-1)}\right]^{Z_0(1)} E\left[s^{I_{ABab}(n-1)}\right]^{Z_{ABab}(1)} s^{Z_{Abab}(1)} \left| Z_i\left(0\right) = 1, Z_j\left(0\right) = 0 \text{ for } j \neq i\right.\right] \\ &= E\left[f_{I_0(n-1)}(s)^{Z_0(1)} f_{I_{ABab}(n-1)}(s)^{Z_{ABab}(1)} s^{Z_{Abab}(1)} \left| Z_i\left(0\right) = 1, Z_j\left(0\right) = 0 \text{ for } j \neq i\right.\right] \\ &= F_i\left(f_{I_0(n-1)}\left(s\right), f_{I_{ABab}(n-1)}\left(s\right), s\right) \end{aligned} \tag{11}$$

where we have used the definition from (7) to complete the last line. We can use the result from (11) with equations (8) and (9) to arrive at recursive relationships for the p.g.f.s of $I_0(n)$ and $I_{ABab}(n)$:

$$\begin{aligned} f_{I_0(n)}(s) &= f_{I_0(n-1)}(s)\, G_0\left(f_{I_{ABab}(n-1)}(s)\right) \\ f_{I_{ABab}(n)}(s) &= G_{ABab}\left(\frac{1}{2}\left(rs + (1-r)\, f_{I_{ABab}(n-1)}(s) + 1\right)\right) \end{aligned} \tag{12}$$

which can be calculated for all n using the boundary conditions $f_{I_{ABab}(0)}(s) = f_{I_0(0)}(s) = 1$, since the total number of type-*Abab* individuals produced at time zero is zero, and consequently the generating functions go to one. Observe that the probability than an introgression event occurs after some time is the probability that all type-*Abab* individual lineages initiated at or before that time become extinct.

Since we start with a single type-0 individual, we can then write the following:

$$P(T > n) = E\left[q^{I_0(n)}\right] = f_{I_0(n)}(q). \tag{13}$$

The hazard rate now follows from combining (13), (12) and (1), which gives the following expression:

$$H(n) = 1 - G_0\left(f_{I_{ABab}(n-1)}(q)\right) \tag{14}$$

which can be calculated for all n using the last equation from (12). Since our hybridization rates are Poisson-distributed with mean m, the p.g.f. in (14) takes a form which gives the following hazard rate:

$$H(n) = 1 - e^{-m\left(1 - f_{I_{ABab}(n-1)}(q)\right)}. \tag{15}$$

And since the offspring distribution of type-*ABab* individuals is Poisson-distributed, we can write the second equation of (12) as follows:

$$\begin{aligned} f_{I_{ABab}(n)}(s) &= e^{-\frac{1}{2}m_{ABab}\left(1-(1-r)f_{I_{ABab}(n-1)}(s)-rs\right)} \\ &= e^{-w_{ABab}\left(1-(1-r)f_{I_{ABab}(n-1)}(s)-rs\right)} \end{aligned} \tag{16}$$

where m_{ABab} represents the expected number of offspring of a single type-*ABab* individual. We take $w_{ABab} = \frac{1}{2}m_{ABab}$ as our fitness measure, since in a stable sexually reproducing population, each individual produces an average of two offspring. Writing $\lim_{n\to\infty} f_{I_{ABab}(n-1)}(q) = \lim_{n\to\infty} f_{I_{ABab}(n-1)}(q) = \hat{f}_{I_{ABab}}(q)$ and $\lim_{n\to\infty} H(n) = \hat{H}(q)$ in (15) and (16) gives the required result for the asymptotic hazard rate.

References

1. Levin, D.A., Francisco-Ortega, J., Jansen, R.K. 1996 Hybridization and the extinction of rare plant species. Conserv. Biol. 10, 10-16 (DOI 10.1046/j.1523-1739.1996.10010010.x)

2. Ellstrand, N.C., Prentice, H.C., Hancock, J.F. 1999. Gene flow and introgression from domesticated plants into their wild relatives. Annu. Rev. Ecol. Systemat. 30, 539-563 (DOI 10.1146/annurev.ecolsys.30.1.539)

3. Hall, R.J., Hastings, A., Ayres, D.R. 2006 Explaining the explosion: modelling hybrid invasions. Proc. R. Soc. 273, 1385-1389 (DOI 10.1098/rspb.2006.3473)

4. Demon, I., Haccou, P., van den Bosch, F., 2007 Introgression of resistance genes between populations: A model study of insecticide resistance in Bemisia tabaci. Theor. Popul. Biol. 72, 292-304 (DOI 10.1016/j.tpb.2007.06.005)

5. Lambert, A., 2006 Probability under weak selection: A branching process unifying approach. Theor. Popul. Biol. 69 419-441 (DOI 10.1016/j.tpb.2006.01.002)

6. Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics 47 713-719

7. Ghosh, A., Haccou, P. 2010 Quantifying stochastic introgression processes with hazard rates. Theor. Popul. Biol. 77, 171-180 (DOI 10.1016/j.tpb.2010.01.002)

8. Ghosh, A., Serra, M.C., Haccou, P., Quantifying time-inhomogeneous stochastic introgression processes with hazard rates. Theor. Popul. Biol. In press (DOI j.tpb.2011.11.006)

9. Stewart Jr., C.N., Halfhill, M.D., Warwich, S.I., 2003 Genetic modification: Transgene introgression from genetically modified crops to their wild relatives. Nature Rev. Genet. 4 806-817 (DOI 10.1038/nrg1179)

10. Kwit, C., Moon, H.S., Warwick, S.I., Stewart Jr., C. N., 2011 Transgene introgression in crop relative: molecular evidence and mitigation strategies. Trends Biotechnol. 29 284-293 (DOI 10.1016/j.tibtech.2011.02.003)

11. Thompson, C.J, Thompson, B.J.P., Ades, P.K., Cousens, R., Carinier-Gere, P., Landman, K., Newbigin, E., Burgman, M.A., 2003 Model-based analysis of the likelihood of gene introgression from genetically modified crops into wild relatives. Ecol. Model. 162 199-209. (DOI 10.1016/S0304-3800(02)00347-2)

12. Kalbfleisch, J.D., Prentice, R.L. 2002 The statistical analysis of failure time data, 2nd edn. John Wiley & Sons. 13. Wei, W., Krone, S.M., 2005 Spatial invasion by a mutant pathogen. J. Theor. Biol. 236 335-348 (DOI 10.1016/j.jtbi.2005.03.016)

14. Parham, P.E., Michael, E., 2011 Outbreak properties of epidemic models: The roles of temporal forcing and stochasticity on pathogen invasion dynamics. J. Theor. Biol. 271 1-9 (DOI 10.1016/j.jtbi.2010.11.015)

15. Zalba, S.M., Songlioni, M.I., Compagnoni, C.A., Belenguer, C.J., 1998 Using a habitat model to assess the risk of invasion by an exotic plant. Biol. Cons. 93 203-208 (DOI 10.1016/S0006-3207(99)00146-9)

16. Michor, F., Nowak, M.A., Iwasa, Y., 2005 Stochastic dynamics of metastasis formation. J. Theor. Biol. 240 521-530 (DOI 10.1016/j.jtbi.2005.10.021)

17. Haccou, P., Jagers, P., Vatutin, V.A. 2005 Branching processes. Variation, growth and extinction of populations. Cambridge University Press.

18. Haccou, P., Meelis, E., 1994 Statistical analysis of behavioural data. Oxford University Press.

19. Eshel, I., 1984 On the survival probability of a slightly advantageous mutant gene in a multitype population: a multidimensional branching process model. J. Math. Biol. 19 201-209

20. Knolle, H., 2004 A discrete branching process model for the spread of HIV via steady sexual partnerships. J. Math. Biol. 48 423-443 (DOI 10.1007/s00285-003-0241-7)

21. Otto, S. P., Whitlock, M. C., 2008. Fixation probabilities and times. In Encyclopedia of Life Sciences. John Wiley & Sons, Ltd.

22. Meirmans, P.G., Bousquet, J., Isabel, N. 2009. A metapopulation model for the introgression from genetically modified plants into their wild relatives. Evol. Appl. 2, 160-171 (DOI 10.1111/j.1752-4571.2008.00050.x)

23. Hanski, I. 1999. Metapopulation Ecology. Oxford University Press, Oxford.