# Calculating hazard rates of introgression with branching processes

Ghosh, A.

# Calculating Hazard Rates of Introgression with Branching Processes

Atiyo Ghosh

Calculating Hazard Rates of Introgression with Branching Processes Atiyo Ghosh

# Promotiecommissie

| | |
|---|---|
| Promotores: | Prof. dr. J.A.J. Metz (Universiteit Leiden)<br>Prof. dr. S.M. Verduyn Lunel (Universiteit Leiden) |
| Copromotor: | Dr. P. Haccou (Universiteit Leiden) |
| Overige leden: | Prof. dr. C.J. ten Cate (Universiteit Leiden)<br>Prof. dr. Y. Iwasa (University of Fukuoka, Japan)<br>Dr. T.J. de Jong (Universiteit Leiden)<br>Prof. dr. P.H. van Tienderen (Universiteit van Amsterdam) |

This PhD project has been funded by the Ecology Regarding Genetically Modified Organisms (ERGO) program, which is managed by the Earth and Life Sciences Council (ALW) of the Netherlands Organisation for Scientific Research (NWO).

Ghosh, A.

Calculating Hazard Rates of Introgression with Branching Processes

Cover Design and Printing: Optima Grafische Communicatie, Rotterdam.



Chapter 4 is being revised for resubmission

Chapter 5 is to be submitted

# Contents

# QUANTIFYING INTROGRESSION RISKS WITH HAZARD RATES

## Introduction

Introgression is the permanent incorporation of genes from the genome of one population into that of another through hybridisation and backcrossing. With the advent of genetically modified (G.M.) crops, the potential that such transgenes might introgress from cultivated to wild plants has become a point of scientific scrutiny. Ellstrand et al. [1999] found that 12 out of the 13 most important food crops in the world hybridise readily with wild relatives, which suggests that the potential for introgression is certainly there. There are multiple and diverse consequences of such gene flow, from the genetic assimilation of crop genes and resulting loss of wild biodiversity [Haygood et al., 2003], to the creation of herbicide-resistant weeds [Owen and Zelaya, 2005]. The discussion of the possible consequences of introgression is beyond the scope of this work. Instead I would like to consider how one should answer an equally fundamental problem of introgression: what is the chance that it occurs, and if it does occur, when is this most likely to happen? This thesis provides a theoretical framework which helps to answer these questions.

Previous model-based studies of introgression are largely deterministic. The number of hybrids appearing in a wild population might initially be small. Consequently, stochasticity in the number of offspring of invaders (demographic stochasticity) is a crucial aspect of introgression, which has been neglected in much previous work [Davis et. al, 1999, Thompson et al., 2003]. Also, little work has been done on defining when introgression actually occurs. For example, Thompson et al. [2003] investigates the number of individuals carrying a foreign gene, and considers that introgression has occurred when this number crosses some threshold. The choice of this level is arbitrary, which makes this approach unsatisfactory.

A quick thought experiment suggests that introgression is inevitable if there is a recurrent gene flow of some fitness-enhancing gene from crops into wild populations. If a single invader has some probability (however small) of initiating a successful invasion into a population, then a flow will almost surely result in a successful lineage becoming established eventually. This is not to say that the risk of introgression is high—even though a successful invasion will eventually occur, it might still take a long time, so the risk of introgression might be acceptably small. Thus, the key to calculating introgression risks lies first in finding the distribution of times at which invasions occur. After this is done, an appropriate measure of introgression risk can be calculated.

A quantity known as the hazard rate is a strong candidate for such a risk measure. The hazard rate of an event is defined as the probability per time unit that the event occurs given that it has not previously occurred. Aside from being an appropriate quantitative measure, hazard rates also provide an intuitive

basis for understanding the risk of events. For example, the hazard rate of an individual winning the weekly lottery is simply the probability per week of winning the lottery (given that the individual has not won the lottery before). Note that the hazard rate can change in time, e.g. a lottery player might buy multiple tickets during some weeks, but buy none at others. Hazard rates have been used widely in the field of survival analysis [Kalbfleisch and Prentice, 2002, for example]. Once the hazard rate has been calculated, then factors affecting introgression risk, and effective risk mitigation strategies can be devised. The hazard rate can be calculated from stochastic models of the introgression process. In this thesis, we use branching process models to do so. For the interested reader, Haccou et al. [2005] provides an introduction to the use of branching process in biology.

The overall aim of quantifying introgression risk must involve combining appropriate theoretical and experimental procedures. The theoretical work within this thesis should be seen as one of three sub-projects with a common goal. The aim of the overall project is to use the carrot (*Daucus carota*) as a case-study in developing a methodology to quantify introgression risk. One sister sub-project investigates short term introgression incidence by crossing several cultivars and wild plants. The results from these crossing experiments can be combined with the results contained within this thesis to estimate introgression risks. Long term introgression incidence is studied by another sub-project, which uses molecular markers to estimate past levels of introgression. The results from studying the molecular markers can be used to validate the predictions from the crossing experiments and theoretical approaches. The whole project benefits from an interdisciplinary collaboration between several members from the Institutes of Biology (IBL) and Environmental Sciences (CML) of Leiden University, and is funded by the Netherlands Organisation for Scientific Research (NWO) as part of their research program 'Ecology Regarding Genetically Modified Organisms' (ERGO).

One of the characteristics of introgression is the occurrence of repeated invasions, each with a small probability of success. Furthermore, changes in fitness of individuals carrying the invading gene may occur, due to changes in the genetic background of the gene. Similar processes occur in many other contexts. For example, the spread of invasive species into new territories, or the invasion of a disease from one host species to another. Consequently, I hope the methodologies and results which follow prove of use to researchers across a range of fields.

**Chapter 1** introduces the general methodology for calculating hazard rates of introgression using a time-homogeneous model of monocarpic perennials with an age-structure. The paper explains how deterministic methods cannot be used to calculate hazard rates. In addition, it proves the monotonic increase of the hazard rate with time for all time homogeneous models, investigates the effect of variance on invasion risk, finds that the hazard rate can either increase or decrease with flowering probabilities, and shows how Taylor approximations of branching processes can lead to biologically plausible arguments.

**Chapter 2** uses a special case from chapter 1 in combination with preliminary results from empirical studies to investigate the hazard rate of introgression from cultivated carrots into their wild relatives. A sensitivity analysis of key life-history parameters is performed, and the results are explained in terms of assumptions and results from chapter 1.

**Chapter 3** investigates the effects of including deterministically varying environments in introgression. Deterministic changes could be human-mediated and used in management strategies for risk mitigation. In particular, it focusses on deterministically varying hybridisation rates on the hazard rate, in scenarios such as crop rotation. In such models, the hazard rate changes with time. Procedures for finding a constant hazard rate which approximates the time-changing changing hazard rate are given. Also, chapter 1 assumes that all backcrosses have identical life-history parameters, and these assumptions are relaxed in chapter 2.

**Chapter 4** presents procedures for incorporating multiple loci and alleles into hazard rate calculations. It shows how the linkage of a transgene to some quantitative trait locus can affect the hazard rate. In addition, it is shown how to calculate hazard rates using computer simulations as well as from branching processes. In previous chapters it was necessary to assume that introgression was occurring into a large wild population to maintain mathematical tractability, but the question remained as to how large was large enough. The use of computer simulations allows us to answer this question, and we find that branching processes do indeed provide a good basis for calculating invasion risks at ecologically realistic population sizes.

**Chapter 5** extends the methodology to include introgression in random environments. In such scenarios, as in chapter 2, the hazard rate conditioned on the environment changes with time. It might be tempting to take the mean hazard rate as a risk measure, but this chapter shows that this would mean that introgression risks might be grossly underestimated at some times. Contrary to previous studies, we find that randomly changing environments can either increase or decrease introgression risks when compared to predictions from models with time-homogeneous parameters.

## References

Davis, S.A., Catchpole, E.A., Pech, R.P., 1999. Models for the introgression of a transgene into a wild population within a stochastic environment, with applications to pest control. Ecol. Model. 119, 267-275.

Ellstrand, N.C., Prentice, H.C., Hancock, J.F., 1999. Gene flow and introgression from domesticated plants into their wild relatives. Annu. Rev. Ecol. Syst. 30, 539-63.

Haygood, R., Ives, A.R., Andow, D.A., 2003. Consequences of recurrent gene flow from crops to wild relatives. Proc. R. Soc. Lond. B 270, 1879-1886.

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

Kalbfleisch, J.D., Prentice, R.D., 2002. The statistical analysis of failure time data, 2nd ed. John Wiley & Sons, New York.

Owen, M.D.K., Zelaya, I.A., 2005. Herbicide-resistant crops and weed resistance to herbicides. Pest Manag. Sci. 61, 301-311.

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.

# CHAPTER 1: QUANTIFYING STOCHASTIC INTROGRESSION PROCESSES WITH HAZARD RATES

Reprinted with minor edits from Ghosh, Haccou, 2010. Theor. Popul. Biol., 77, 171-180

## Abstract

Introgression is the permanent incorporation of genes from one population into another through hybridization and backcrossing. It can have large environmental consequences, such as the spread of insecticide or herbicide resistant genes, the escape of transgenes from genetically modified crops, and the invasion of exotic genes into new habitats. Introgression usually involves strong random components, such as rare hybridization and backcrossing events, and demographic variation in reproduction and survival. Most introgression studies ignore these random effects, and consequently fail to accurately assess the risk of introgression. This paper presents a methodology for quantifying stochastic introgression processes, based on multitype branching process models. We derive a quantity called the hazard rate, which can be used to investigate how the risk of introgression depends on crop management and life history.

## 1. Introduction

Introgression is the permanent incorporation of genes from one population into another, through hybridization and backcrossing (Riesberg and Wendel, 1993; Ellstrand et al., 1999; Hails and Morley, 2005). This may result in the spread of insecticide or herbicide resistant genes (Snow et al., 1999; Demon et al., 2007), the escape of transgenes from genetically modified crops (Rieger et al., 2002; Reichman et al., 2006), or the incorporation of genes from exotic species into genomes of local species (Huxel, 1999; Allendorf et al., 2001; Abbott et al., 2003). The potential environmental effects of introgression are severe. For instance, there are serious concerns that transgene escape might produce robust weeds (Maan, 1987; Snow et al., 1999; Thompson et al., 2003; Kelly et al., 2005) that can outcompete other species and reduce biodiversity (Levin et al., 1996; Jenczewski et al., 2003; Ellstrand, 2003).

Mathematical models provide a means to study the likelihood of introgression given certain environmental and species-specific conditions. The great advantage of models is that they allow us to perform experiments that are either too dangerous, impractical, or simply impossible to carry out empirically. Furthermore, models can pinpoint which parameters are crucial for introgression risk, and whose values should therefore be determined empirically.

Introgression processes contain strong stochastic components. Hybridization and backcrossing events usually occur at a very low rate. Hybrids and initial backcrosses are often less viable and fertile than the wild type (Hauser et al., 1998), since their genetic backgrounds are adapted to different conditions. Therefore, initial hybrid populations are usually small, and highly affected by demographic

stochasticity. As a consequence, even if foreign genes provide a fitness advantage in later backcrosses, it will usually take several invasions before they are established permanently. The number of individuals carrying foreign alleles will be highly variable, and rise and fall during the initial stages of introgression. The use of deterministic models, that fail to take such key features of introgression into account, can be very misleading.

Nevertheless, stochastic models are seldom used in introgression studies. Exceptions are Davis et al. (1999) and Thompson et al. (2003), who considered the effects of stochastic environmental variation, but ignored demographic stochasticity. Haygood et al. (2003) studied the conditions under which transgenes can become fixed due to genetic drift in small populations with repeated invasions. Haygood et al. (2004) examined the repeated invasions of a transgene with a small fitness advantage.

For practical applications, general quantitative measures that characterize stochastic introgression processes are indispensable. The above-mentioned studies do not provide these, since they are all based on simulations. Our aim is to develop such measures.

We consider a situation where foreign genes invade repeatedly into a resident population, and each invasion has a small probability of leading to a permanent lineage. This is similar to the case studied by Haygood et al. (2004), but we consider a more general model, and our main results are not based on simulations. It is obvious that under these conditions a permanent introgressed lineage will be founded eventually. Before such an introgression event happens, however, there can be an extensive period of failed invasions. Introgression risk is largely determined by the duration of this period. We derive a measure that quantifies the distribution of these lengths in an intuitive way, called the hazard rate.

The hazard rate is the probability per time unit that a random event occurs, given that it has not happened before. It quantifies how the instantaneous risk of introgression events changes in time. For instance, how quickly this risk increases after cultivation of a transgene crop has started, or decreases after such cultivation is stopped, and how this relates to life history characteristics of the crop and its wild relatives. Studies of hazard rates can indicate which periods are especially risky, and thus provide information for optimizing monitoring and management programs.

Hazard rates were first used in medical statistics, to analyze mortality risks (Kalbfleisch and Prentice, 2002). For several decades they have also been applied in behavioral data analysis (Haccou and Meelis, 1994). They have not been used in introgression studies before now.

We derive the hazard rate from a multitype branching process model of hybrid population dynamics (e.g., Haccou et al., 2005, section 2.2). Demon et al. (2007) used such a model to study the effects of whitefly life history parameters on the establishment probability of an insecticide resistance gene, but they only considered a single introduction. Serra and Haccou (2007) were the first to calculate hazard rates from a branching process model. They assumed that individuals that have on average less than one offspring (the so-called subcritical type) can produce mutants with an average of more than one offspring (the supercritical type). Escape from extinction occurs as soon as a supercritical individual that initiates

a permanent lineage is produced. In this paper we generalize their method to situations with repeated invasion of individuals of the subcritical type, population structure, and density-dependent competition with an established resident type.

As an example, we consider a model for a monocarpic (i.e., a plant which flowers once then dies) monoecious (i.e., flowers have male as well as female functions) non-selfing species. In numerical examples we use parameter ranges that are deemed to be realistic for Daucus carota (the carrot), but we stress that the model is a caricature, and not meant to give a complete description of introgression in this species. We illustrate the set-up of a branching process model and derivation of the hazard rate. Furthermore we give numerical procedures for calculating this function, and analytic approximations.

While we focus on introgression, our approach may be applied in other fields as well, since repeated invasions with fitness bottlenecks occur in many biological contexts. For example, when a virus colonizes a new host species, its initial reproductive ratio will be small, but after a few mutations this can increase. Another example is the initiation and growth of tumors, where cells usually have to go through several mutations before they produce a successful lineage (see e.g., Michor et al., 2006).

## 2. The model

We consider a plant species that flowers only once and then dies. To account for an ageing effect, we distinguish one-year old plants from older ones. We assume that there is a large and stable population of wild plants. By pollen flow from a neighboring crop, stochastic numbers of hybrid seeds are produced each year. Hybrid production can be followed by repeated backcrossing with wild plants. Seeds compete to germinate and survive their first year. Hybrid plants are considered to be less fit than wild individuals, whereas backcrossed individuals have some probability of producing a permanent introgressed lineage. Individuals of the first backcross (BC1) and further backcrosses are assumed to have identical life history parameters, and are therefore not distinguished.

In summary, there are six types of plants in the model: one-year old plants (types 1, 3, 5), and plants that are two or more years old (types 2, 4, 6); and in addition to being characterized by their age, plants can also be either wild (types 1 and 2), hybrids (3, 4), or backcrossed (5, 6).

Fig. 1 illustrates the introgression scheme and the life history.

2.1. **Dynamics of the wild population.** Since it is assumed that the wild population is large, we use a deterministic model for its dynamics:

$$\begin{pmatrix} z_1(t+1) \\ z_2(t+1) \end{pmatrix} = \begin{pmatrix} p_0(z_1(t), z_2(t)) r_1 m_1 & p_0(z_1(t), z_2(t)) r_2 m_2 \\ p_1(1-r_1) & p_2(1-r_2) \end{pmatrix} \begin{pmatrix} z_1(t) \\ z_2(t) \end{pmatrix} \quad (1)$$

where $z_i(t)$ represents the population density (number per unit area) of type-i plants in year t, $r_i$ its flowering probability, $m_i$ the average number of seeds it produces and $p_i$ the probability that it survives one year if it does not flower. Furthermore, we assume that there is a density-dependent probability that seeds germinate and survive their first year, $p_0(z_1(t), z_2(t))$.

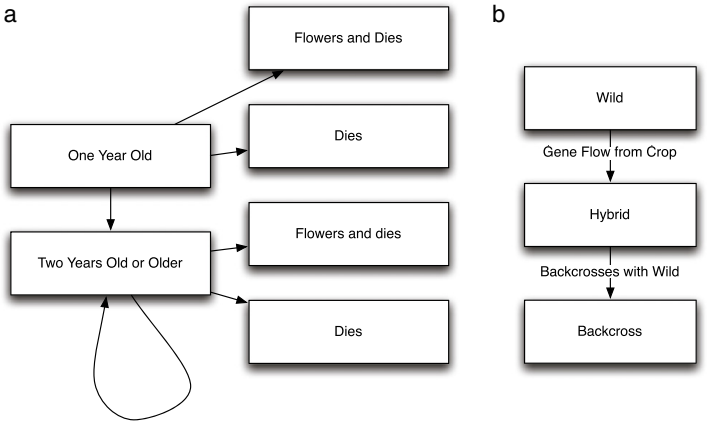


FIGURE 1. (a) Schematic representation of the life history used in the model. (b) A schematic representation of the process by which an introgressed gene moves through a population.

This model has one stable and positive equilibrium. At this equilibrium we have:

$$p_0\left(\hat{z}_1, \hat{z}_2\right) = \frac{1 - p_2\left(1 - r_2\right)}{m_1 r_1\left(1 - p_2\left(1 - r_2\right)\right) + p_1 m_2 r_2\left(1 - r_1\right)} \tag{2}$$

We will denote this value by $p_0$, and refer to it as the germination probability.

Furthermore, we define:

$$\zeta_{WT} = \hat{p}_0^{-1} = r_1 m_1 + \frac{r_2 m_2\left(1 - r_1\right) p_1}{1 - p_2\left(1 - r_2\right)} \tag{3}$$

which represents the expected number of seeds produced by one wild type individual.

2.2. **Invasion dynamics of hybrids and backcrosses.** Because the population of wild plants is large and the numbers of hybrid and backcrossed individuals are initially small, it can be assumed that these individuals do not interact with each other, but only with wild plants. This has several implications. Firstly, since we consider a non-selfing species, reproduction can only occur through outcrossing with wild plants. Secondly, competition occurs only with the wild population, implying that the seed germination probability equals $\hat{p}_0$. For convenience, we assume that there are no other factors beside this competition that affect germination probability of hybrids and backcrosses. The model can be easily generalized in this respect. Since hybrid and backcrossed plants do not affect each others reproduction and survival initially, their invasion dynamics can be modeled as a branching process.

In the branching process model, flowering individuals of type $i \in \{3, 4, 5, 6\}$ produce a stochastic number of offspring, denoted by $\xi_i$, with expectation $m_i$. The probabilities $r_i$, and $p_i$ are as defined in the previous section, and are assumed to lie between zero and one. The production of hybrid seeds is modeled by means of an artificial type, which we will call type 0. There is one permanently present

individual of type 0 that produces a stochastic number of hybrid seeds, $\xi_0$, according to some probability distribution with expectation $m_0$ in each year. Figure 2 shows a schematic summary of the invasion dynamics.

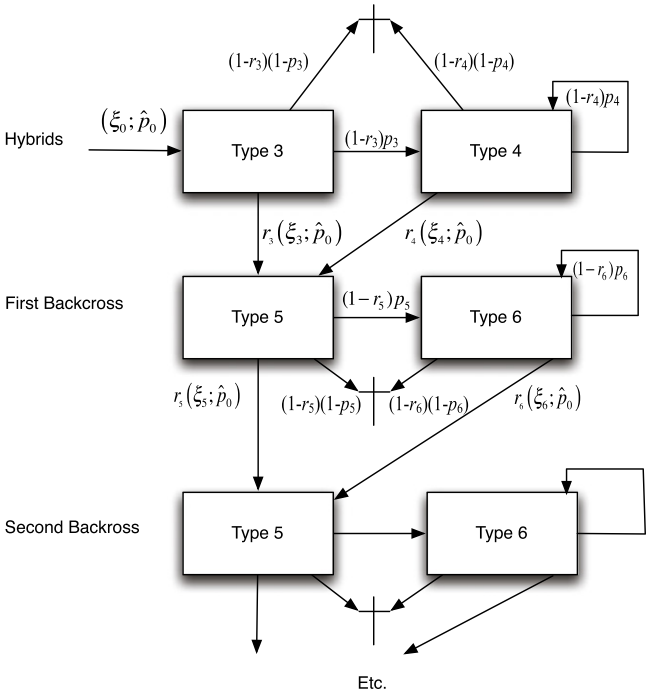


FIGURE 2. Schematic representation of the branching process model for invasion dynamics. $(\xi;\hat{p}_0)$, $(i = 0, 3, 4, 5, 6)$ represents the production of $\xi_i$ seeds, each of which has a germination probability $\hat{p}_0$. Second and further backcrosses are assumed to have identical life history parameters to BC1.

## 3. Stochastic versus deterministic invasion dynamics

Most introgression studies are based on deterministic rather than stochastic invasion models. We here demonstrate the difference between the two approaches. Let $Z_i(t)$ denote the number of type-$i$ individuals in year $t$, then the deterministic equivalent of the branching process model for invasion dynamics is:

$$\begin{pmatrix} Z_3(t+1) \\ Z_4(t+1) \\ Z_5(t+1) \\ Z_6(t+1) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ p_3(1-r_3) & p_4(1-r_4) & 0 & 0 \\ \hat{p}_0 r_3 m_3 & \hat{p}_0 r_4 m_4 & \hat{p}_0 r_5 m_5 & \hat{p}_0 r_6 m_6 \\ 0 & 0 & p_5(1-r_5) & p_6(1-r_6) \end{pmatrix} \begin{pmatrix} Z_3(t) \\ Z_4(t) \\ Z_5(t) \\ Z_6(t) \end{pmatrix} + \begin{pmatrix} m_0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{4}$$

In a population starting with at least one $Z_i(t)$ larger than zero and no immigration (if $m_0 = 0$), the numbers of types 5 and 6 will grow if and only if the dominant eigenvalue of the matrix:

$$\begin{pmatrix} \hat{p}_0 r_5 m_5 & \hat{p}_0 r_6 m_6 \\ p_5 (1 - r_5) & p_6 (1 - r_6) \end{pmatrix} \tag{5}$$

is larger than one. This leads to the condition

$$\zeta_{BC} > \zeta_{WT}, \tag{6}$$

where $\zeta_{BC}$ is the expected number of seeds produced by a single backcrossed plant:

$$\zeta_{BC} = r_5 m_5 + \frac{r_6 m_6 (1 - r_5) p_5}{1 - p_6 (1 - r_6)} \tag{7}$$

With immigration ($m_0 > 0$) and initial condition $Z_3(0) = \ldots = Z_6(0) = 0$, the deterministic model then predicts an exponential growth of $Z_5(t)$ and $Z_6(t)$ from year 2 onwards.

According to the stochastic model, exactly the same condition must hold for a process starting with one individual of type 5 to have a positive establishment chance (see section 4.1). In that case, repeated invasions will eventually lead to exponential growth of the numbers of type-5 and 6 individuals. Before that happens, however, there may be several failed invasions, and even production of backcrossed individuals whose lineage fails. Thus, the numbers of individuals that carry the foreign gene can rise and fall, and periods in which the gene is present or absent from the population may alternate, before initiation of a permanent lineage. This is illustrated in Fig. 3, which shows the trajectory predicted by the deterministic model as well as the results of several simulations of the branching process. As can be seen, the length of the initial period of failed invasions is highly variable and may be quite large. Furthermore, the numbers of individuals with the foreign allele at any given time differ strongly between simulations.

## 4. Derivation of the hazard rate

Let $T$ denote the time at which an introgression event occurs, i.e., the time at which the first type-5 individual whose lineage does not die out is produced. In the three examples of Fig. 3, $T$ is respectively about 2, 60, and larger than 100 years. The hazard rate, denoted by $H_n(q)$, equals:

$$H_n(q) = P(T = n \,|T > n - 1) = \frac{P(T > n - 1) - P(T > n)}{P(T > n - 1)} \tag{8}$$

From this equation it can be seen that this function fully characterizes the distribution of $T$.

### 4.1. Establishment probability of an introgressed lineage from one type-5 individual.

We first consider the fate of a lineage that starts with one type-5 individual. According to branching process theory, lineages either go extinct or they grow infinitely large. In situations where backcrosses are more fit than the wild type, the extinction probability of a lineage from one type-5 individual, denoted by $q_5$, is less than one, and the branching process model predicts indeterminate growth with a probability $1 - q_5$. It is unreasonable, however, to assume that populations can grow infinitely large, so we will interpret $1 - q_5$ as the probability

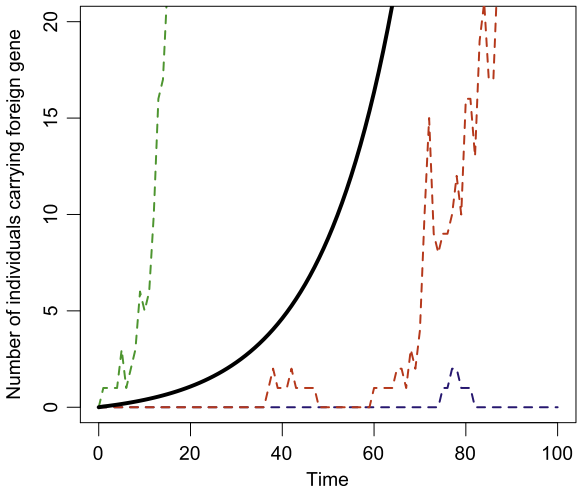


FIGURE 3. Total number of individuals containing the foreign gene, $\sum_{i=3}^{6} Z_i(t)$ according to the deterministic model (solid line), and the three simulation runs of the stochastic model with Poisson-distributed seed productions (dotted lines). Parameter values: $\hat{p}_0 = 0.00079$, $m_0 = 50$, $r_3 = r_4 = r_5 = r_6 = 0.9$, $p_3 = p_4 = p_5 = p_6 = 0.7$, $m_3 = m_4 = 850, m_5 = m_6 = 900$.

that the lineage reaches some size that is large enough to guarantee its permanent presence.

The calculation of extinction probabilities and hazard rate is based on so-called probability generating functions. Let $X$ be a non-negative discrete random variable, then its probability generating function (p.g.f.) is a function from $\mathbb{R}$ to $\mathbb{R}$ , which is defined as $E\left[s^X\right]$, where $E[\ ]$ denotes expectation and $s$ can take any value on the interval $[0, 1]$.

Let $G_i(s)$ be the p.g.f. of the number of seeds produced by such an individual of type $i$ $(i = 0, ..., 6)$, then:

$$\begin{aligned} q_5 =& (1-p_5)(1-r_5) + (1-r_5)p_5q_6 + r_5 \sum_k P(\xi_5 = k) \sum_{l \le k} \binom{k}{l} (1-\hat{p}_0)^{k-l} \hat{p}_0^l q_5^l \\ =& (1-p_5)(1-r_5) + (1-r_5)p_5q_6 + r_5 \sum_k P(\xi_5 = k) \left(\hat{p}_0 q_5 + (1-\hat{p}_0)\right)^k \\ =& (1-p_5)(1-r_5) + (1-r_5)p_5q_6 + r_5 G_5(\hat{p}_0 q_5 + (1-\hat{p}_0)) \end{aligned} \tag{9}$$

The first term represents the probability that a type-5 individual does not flower and does not survive; the second term represents the probability that the individual does not flower, and survives to become a type-6 individual, which initiates a lineage that goes extinct; the last term equals the sum of probabilities that the type-5 individual flowers, produces $k$ seeds of which $l$ germinate and produce type-5 individuals that each produce a lineage which will become extinct. We can write (9) as follows:

$$q_5 = a_5(q_5) + b_5 q_6 \tag{10}$$

where we have introduced

$$a_i(q) = r_i G_i(\hat{p}_0 q + (1-\hat{p}_0)) + (1-r_i)(1-p_i) \quad \text{and} \quad b_i = (1-r_i)\,p_i \tag{11}$$

for $i = (1, \ldots, 6)$. These quantities will appear frequently in subsequent analyses.

In order to solve (10), we must first derive an expression for q6. Analogous to (10) we can derive:

$$q_6 = a_6(q_5) + b_6 q_6. \tag{12}$$

The solution of Eqs (10) and (12) depends on the forms of the functions $G_5(s)$ and $G_6(s)$, as well as the parameter values. For ease of notation, we will denote $q_5$ by $q$ from now on. From (10) and (12) it follows that:

$$q = a_5(q) + \frac{b_5 a_6(q)}{1-b_6} \tag{13}$$

The smallest positive value of $q$ that satisfies (13) represents the extinction probability of a lineage started by one type-5 individual. This value is smaller than one if and only if the derivative of the right-hand side of (13) evaluated at $q = 1$ is larger than one, which leads to inequality (6).

4.2. **Calculation of the hazard rate.** Let $I_i(n)$ ($i \in \{0, 3, 4\}$) be the cumulative number of type-5 individuals produced by type-3 and 4 individuals up to and including time $n$, given that we start with a population of a single type-$i$ individual. We denote the p.g.f. of $I_i(n)$ by $f_{I_i(n)}(s)$. It is easily seen that:

$$P(T > n) = E\left[q^{I_0(n)}\right] = f_{I_0(n)}(q) \tag{14}$$

since $T > n$ implies that all type-5 individuals produced up to and including time n must fail to establish a permanent lineage. In the Appendix it is shown that:

$$\begin{aligned}
f_{I_0(n)}(q) &= f_{I_0(n-1)}(q)\, G_0\left(\hat{p}_0 f_{I_3(n-1)}(q) + (1-\hat{p}_0)\right) \\
f_{I_3(n)}(q) &= r_3 G_3(\hat{p}_0 q + (1-\hat{p}_0)) + (1-r_3)\, p_3 f_{I_4(n-1)}(q) + (1-r_3)(1-p_3) \\
f_{I_4(n)}(q) &= r_4 G_4(\hat{p}_0 q + (1-\hat{p}_0)) + (1-r_4)\, p_4 f_{I_4(n-1)}(q) + (1-r_4)(1-p_4)
\end{aligned} \tag{15}$$

These equations are solved iteratively with initial condition $f_{I_i(0)}(q) = 1$, $i \in \{0, 3, 4\}$ to find $f_{I_0(n)}(q)$. Putting (8), (14) and (15) together gives:

$$H_n(q) = 1 - \frac{f_{I_0(n)}(q)}{f_{I_0(n-1)}(q)} = 1 - G_0\left(\hat{p}_0 f_{I_3(n-1)}(q) + (1-\hat{p}_0)\right) \tag{16}$$

For $n = 0$ or 1, the hazard rate equals 0, since no type-5 individuals can be produced before year 2. For the considered model, it is possible to derive an explicit expression for the hazard rate. For $n \geq 2$, it equals (see Appendix) :

$$H_n(q) = 1 - G_0\left(\hat{p}_0\left(a_3(q) + b_3\left(a_4(q)\,\frac{1-{b_4}^{n-2}}{1-b_4} + {b_4}^{n-2}\right)\right) + (1-\hat{p}_0)\right) \tag{17}$$

and its asymptotic value is:

$$\hat{H}(q) = 1 - G_0\left(\hat{p}_0\left(a_3(q) + \frac{b_3 a_4(q)}{1-b_4}\right) + (1-\hat{p}_0)\right) \tag{18}$$

4.3. **Shape of the hazard rate.** The recurrence relations for $f_{I_3(n)}(q)$ and $f_{I_4(n)}(q)$ in (15) can be considered from a different angle. Consider a multi-type branching process with three types, numbered 3, 4 and 5, and let $Q_i(n)$ be the probability that a process starting with one individual of type $i$ will go extinct at or before time n. Furthermore, assume that $Q_5(n)$ is constant and equal to $q$. Then (15), with $f_{I_i(n)}(q)$ replaced by $Q_i(n)$ $(i = 3, 4)$ specifies recurrence relations for these extinction probabilities, but with different initial conditions: $f_{I_i(0)}(q) = 1$ , whereas $Q_i(0) = 0$. This implies that the $f_{I_i(n)}(q)$ decrease with $n$, whereas the $Q_i(n)$ increase. Of course, it is also evident from their definitions that this should be so. This conclusion is also valid for models describing more complicated life histories.

This equivalence can be used to derive properties of the $f_{I_i(n)}(q)$. From branching process theory (e.g., Athreya and Ney, 1972, chapter 5) it follows that, when $q$ is given, the recurrence relations in (15) have one equilibrium with values in [0,1]. Let $f_{I_i}(q)$ denote the equilibrium values of $f_{I_i(n)}(q)$. If $q = 1$, there is one equilibrium at the point where all $f_{I_i}(q)$ are equal to one. In that case $f_{I_3(n)}(q) = f_{I_4(n)}(q) = 1$ for all $n$, and the hazard rate equals zero.

If $q$ is less than one, there is one equilibrium point at smaller values of $f_{I_i}(q)$, which is stable. Since hybrid individuals have a positive chance of having no surviving offspring, this equilibrium is larger than zero, even when q equals zero. In this case $f_{I_3(n)}(q)$ decreases monotonically to a constant equilibrium value, and it follows from (16) that the hazard rate increases monotonically to an asymptote between zero and one.

A similar expression to (16) will hold for any process where the numbers of hybrids that are introduced each time unit are independent and identically distributed. Therefore, the same conclusion holds for any introgression process with this immigration structure.

When q is close to one, approximations for the extinction probabilities $Q_i$ in slightly supercritical processes, such as given in Haccou et al. (2005, section 5.6) can be used to approximate the $f_{I_i}(q)$. An example of such an approximation is given in section 5.3.

## 5. Results

5.1. **The shape of the hazard rate and distribution of T.** In the initial numerical analyses (Fig. 4) we assume that the seed production distributions of types 3 and 4 are either Geometric ($G_i(s) = 1/(1+(1-s)m_i)$ ) or Poisson ($G_i(s) = e^{-m_i(1-s)}$), and that the numbers of hybrid seeds produced per generation are Poisson-distributed.

As shown in the previous section, the hazard rate is zero in the first year and then increases monotonically in time, to an asymptotic level smaller than one. In this example, the asymptotic value is reached rapidly (see Fig. 4a). If the hazard rate would be constant, the times $T$ would have a Geometric distribution (see e.g., Feller, 1968, section 13.9). The results indicate that we can well approximate the distribution of $T$ with a time-lagged Geometric distribution:

$$P(T=t) \approx P(T_g+1=t) = \left(1-\hat{H}(q)\right)^{t-2}\hat{H}(q), \text{ for } t \geq 2 \qquad (19)$$
$$P(T=t) = 0, \text{ for } 0 \leq t < 2,$$

where $T_g$ represents a Geometrically distributed random variable. Numerical simulations, as shown in Fig. 4b, support the effectiveness of this approximation.

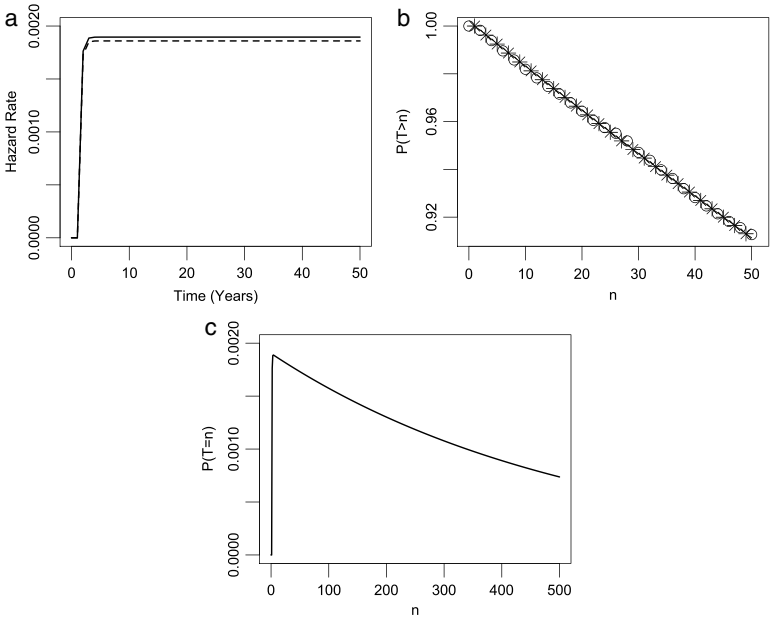


FIGURE 4. (a) The time dependence of the hazard rate (c.f. Eq. (17)) for Poisson (solid line) and Geometrically (dotted line) distributed seed production. Hybrids are formed according to a Poisson distribution. (b) The survival function of $T$ for Poisson-distributed seed numbers (line). Numerical simulations of 10,000 iterations (circles) and the approximation to the survival function of $T$ (stars) with a time-lagged Geometric distribution. (c) The exact distribution of $T$ for Poisson-distributed seed production. In all the three figures the parameter values are: $m_0 = 50$, $m_3 = m_4 = 800$, $\hat{p}_0 = 0.001$, $p_3 = p_4 = 0.7$, $r_3 = r_4 = 0.9$, $q = 0.95$.

As a consequence, a good approximation for the expectation of T is $1+1/\hat{H}(q)$. Note, however, that the distribution of $T$ is very skewed (see Fig. 4c). This implies that most introgression events will occur before the expected time. For example, with the parameters used in Fig. 4, the expectation of $T$ equals 529 years, whereas half of the introgression events occur at or before 368 years.

5.2. **Effects of hybrid survival and flowering probabilities, hybrid seed production, and hybrid formation.** It can be shown straightforwardly that

the asymptotic hazard rate is monotonically non-decreasing with $r_4$, $p_3$, and $p_4$ (see Appendix). Numerical examples are given in Fig. 5.

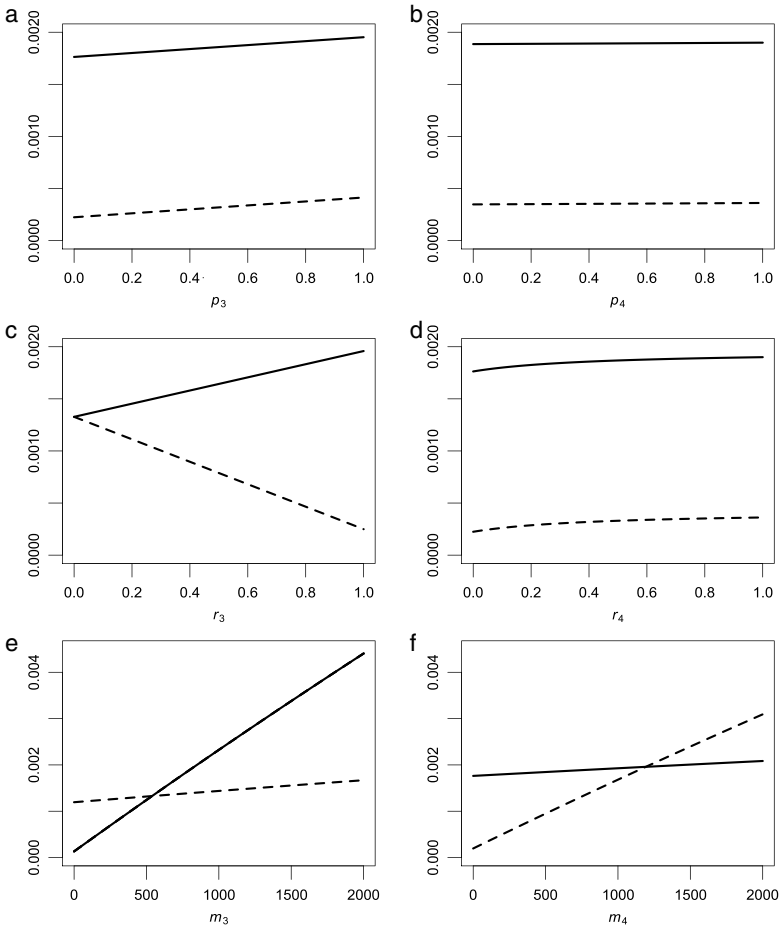


FIGURE 5. Dependence of the asymptotic hazard rate (c.f. Eq. (18)) on parameters $p_3$ (a), $p_4$ (b), $r_3$ (c), $r_4$ (d), $m_3$ (e) and $m_4$ (f). Default parameter values are $m_0 = 50$, $m_3 = 800$, $m_4 = 800$, $\hat{p}_0 = 0.001$, $p_3 = 0.7$, $p_4 = 0.7$, $r_3 = 0.9$, $r_4 = 0.9$ , $q = 0.95$, except in dottedlines in (a), (b), (c) and (d) where $m_3 = 100$, and dotted lines in (e) and (f) where $r_3 = 0.1$.

Since $a_i(q)$ decreases in $m_i$ and $G_0(s)$ increases in $s$, it follows from (18) that the asymptotic hazard rate increases monotonically with $m_3$ and $m_4$. The rate of this increase is linked to $r_3$, as illustrated in Figs. 5e and f. If $r_3$ is low, a smaller number of type-3 individuals flower, so $m_3$ has a smaller effect on the asymptotic hazard rate. At the same time, more type-4 individuals are produced, so a decrease in $r_3$ is associated with an increase in the sensitivity of the asymptotic hazard rate to $m_4$.

Note that for Poisson and Geometric distributions,

$$\lim_{m_i \to \infty} G_i(s) = 0 \text{ if } s \neq 1. \tag{20}$$

Thus, when we let $m_3$ and $m_4$ tend to infinity, the right hand side of (18) tends to:

$$1 - G_0\left((1-\hat{p}_0) + \hat{p}_0\left((1-r_3)(1-p_3) + \frac{b_3(1-r_4)(1-p_4)}{1-b_4}\right)\right) \tag{21}$$

where the argument of $G_0$ can be interpreted as the probability that a hybrid seed produces a non-flowering plant, or does not germinate. This result implies that at high hybrid fecundities, the production of hybrids becomes the limiting factor of introgression.

To further examine the effects of the shapes of seed number distributions on the asymptotic hazard rate, we studied its Taylor approximation in the vicinity of $q = 1$ (derivation see Appendix):

$$\hat{H}(q) = \begin{pmatrix} \hat{p}_0^2 m_0\left(r_3 m_3 + \frac{b_3}{1-b_4} r_4 m_4\right)(1-q) \\ -\frac{1}{2}\beta_0 \hat{p}_0^4 \left(r_3 m_3 + \frac{b_3}{1-b_4} r_4 m_4\right)^2 (1-q)^2 \\ -\frac{1}{2} m_0 \hat{p}_0^3 \left(r_3\beta_3 + \frac{b_3}{1-b_4} r_4 \beta_4\right)(1-q)^2 + O\left((1-q)^3\right) \end{pmatrix} \tag{22}$$

where $\beta_i = G''_i(1) = E[X_i(X_i-1)] = Var[X_i] + E[X_i](E[X_i]-1)$, $X_i$ represents the number of seeds produced by a type-$i$ individual and $Var[X_i]$ represents its variance. This result indicates that the hazard rate decreases with increasing variance of seed production by hybrids.

The direction of the effect of $r_3$ depends on the values of other parameters (see Fig. 5c). To get an indication of the parameter ranges where this changes, we studied the first-order term of the Taylor approximation in (22). This term increases in $r_3$ if the expected number of seeds produced by a flowering one-year old hybrid plant is larger than the expected number of seeds it will produce if it postpones flowering:

$$m_3 > \frac{p_3 r_4 m_4}{1-(1-r_4)p_4}. \tag{23}$$

Numerical work confirmed that this inequality provides a good indicator of the switching boundary.

There will be a steady increase in the asymptotic hazard rate with increasing $m_0$, unless $q = 1$. This can be seen from (18) and (22). It follows from (20) that for Poisson-distributed hybrid formation, the asymptotic hazard rate approaches one at large $m_0$. The approximation in (22) indicates that an increased variance in the number of hybrids produced results in a lower asymptotic hazard rate. This can also be seen in Fig. 4a, since a Geometric distribution has a larger variance than a Poisson distribution with the same expectation (see also section 5.4).

5.3. **Effects of backcross fitness.** To examine the effects of backcross fitness relative to the wild type, on introgression success probability, we use the approximation for establishment success of a slightly supercritical branching process, which was derived by Haldane (1927) and later by Eshel (1981) in a more general setting (see also section 4.3). This approximation is based on the second order

Taylor approximation of the right-hand side of (13) in the point $\zeta_{WT} = \zeta_{BC}$, and leads to:

$$(1-q) \approx \frac{2\left(\frac{\zeta_{BC}}{\zeta_{WT}} - 1\right)\zeta_{WT}^2}{\eta_{BC}} \tag{24}$$

where $\eta_{BC} = a''_5(1) + \frac{b_5}{1-b_6} a''_6(1)$. Since $\eta_{BC}$ increases with increasing variance of backcrossed seed production, such variance decreases establishment success, and therefore decreases the hazard rate. Substituting the approximation in (24) in the Taylor-approximation of the asymptotic hazard rate in (22) gives an approximation of the asymptotic hazard rate for situations where the fitnesses of the backcrosses and wild type are nearly equal. Since the variance in the backcross seed production appears in the first order term of the resulting approximation, it has a larger effect on the hazard rate than the variances of hybrid seed production, or hybrid formation, which only affect the second order terms. This is illustrated in Fig. 6.

5.4. **Effects of variances.** The Taylor approximations of the asymptotic hazard rate presented in sections 5.2 and 5.3 indicate that variance in seed production and hybrid formation reduces the hazard rate. These effects cannot be studied well, however, with the models that we used up to now, because in Poisson and Geometric distributions the variance and the mean are interdependent. Furthermore, in the previous examples we used relatively large values for the mean seed productions, $m_i$. To study the effects of variance more closely, we use a model with lower mean values, and so-called Linear fractional distributions for seed production and hybrid formation. The probability generating functions for Linear fractional distributions have the form: $G_i(s) = 1 - (b/(1-c)) + (bs/(1-cs))$, with $c \in (0,1)$ and $b \in (0, 1-c)$. The mean and variance, are, respectively: $b/(1-c)^2$ and $\left(b(1-c)^2 - b^2\right)/(1-c)^4$.

As can be seen from Fig. 6, increasing variances can indeed reduce the hazard rate considerably, and changes in variance in backcross seed production have the largest effect. Both results agree with the predictions from the Taylor approximations.

## 6. Discussion

In this paper we showed how to model and quantify stochastic introgression processes. As illustrated in section 3, predictions of stochastic introgression models differ strongly from those of their deterministic analogues (Fig. 3). Most importantly, deterministic models ignore the initial period before an introgression event occurs. This period, however, strongly determines the risk of introgression, since exponential growth only occurs after initiation of a successful lineage. Furthermore, the high variance between the results of different simulation runs indicates that a representation of population sizes by their expectations, as in deterministic models, is not very useful.

Whereas we focused on the distribution of the lengths of initial periods, previous introgression research, based on deterministic models, considered changes in population sizes of individuals carrying the foreign gene. A complete characterization of the process would involve both aspects. Serra (2006) showed that for a two-type model with one supercritical type, the distribution of the time until the supercritical population reaches a high, fixed level $x$ can be approximated well by

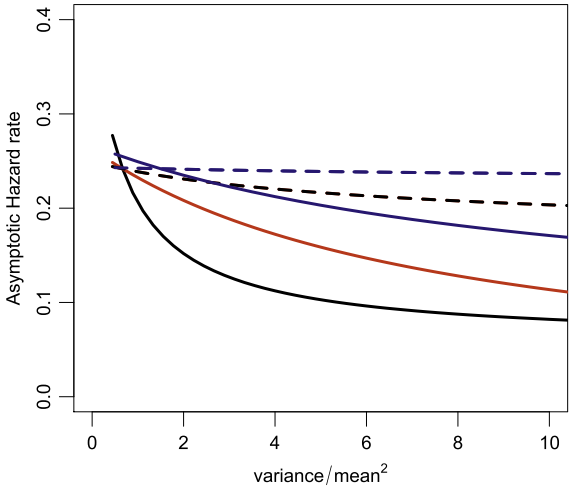


FIGURE 6. The asymptotic hazard rate (Eq. (18)) plotted against the variance of backcrossed seed production (black), hybrid seed production (blue), and hybrid formation (red). Default parameter values: $\hat{p}_0 = 0.8$, $p_3 = p_4 = p_5 = p_6 = 0.7$, $r_3 = r_4 = r_5 = r_6 = 0.9$, $m_0 = m_5 = m_6 = 1.5$, $m_3 = m_4 = 1.0$. All distributions are Linear Fractional, with default variance 1.5. Solid lines represent cases where one-year old plants variances are changing; dotted lines represent the case for changing the variance of older plants. See main text for further details.

the distribution of the time $T$ until an introgression event occurs plus the time it takes a single type supercritical branching process starting with one individual to grow up to level $x$. The simulation results presented in Fig. 3 indicate that a similar result holds in this case. This is a subject of further research.

We introduced the hazard rate as a measure to quantify the distribution of $T$. This function specifies the instantaneous risk of an introgression event in the course of time. It is comparable to e.g., the age-dependent mortality risk of humans in demography. We showed that when there is an initial fitness bottleneck and repeated hybrid formation, the hazard rate is initially zero and then increases to an asymptote between zero and one. In the presented example, this increase happens very quickly, and the distribution of times until introgression events is well approximated by a Geometric distribution with a time lag. In situations with, e.g., multi-generation bottlenecks, or the presence of a seed bank with a gradual build-up of introgressed seeds, the hazard rate will increases to its asymptotic level more gradually. We expect, however, that time-lagged Geometric distributions will still provide good approximations. This implies that after the initial lag time, the process can be considered as a lottery where each year there is a constant chance $\hat{H}(q)$ of an introgression event. As a consequence, the distribution of $T$ will be very skewed (as in, e.g., Fig. 5), and the expected value of $T$ will be a misleading

measure of introgression risk, since in the majority of cases introgression events will take place before this time.

In the relatively simple example considered here, it is possible to derive an explicit expression for the hazard rate. For more complicated cases this will usually not be possible. The values of $q$ and $f_{I_0(n)}(q)$ then have to be determined numerically. Note that this is a far more efficient procedure than performing simulations to estimate the hazard or survival function of $T$ and, furthermore, it provides exact values (apart from numerical errors, which can be minimized by algorithm optimization) rather than estimates. This is especially important in the tail of the distribution of $T$ (i.e., for large values of $n$), where the precision of estimators is low. Furthermore, whereas explicit expressions may not always be available, recurrence relations such as (15) can also be used to derive approximations, or perform analytical studies. Procedures for approximating extinction probabilities of multitype processes can be applied for that purpose, by making use of the analogy explained in section 4.3.

As an example, we studied a simple model of a monocarpic plant species. We found that the asymptotic hazard rate increases monotonically in all life history parameters except the first-year flowering probability, $r_3$ (Fig. 5). Effects of $r_3$ reflect the trade offs between seed production in the first or later years: the hazard rate decreases with $r_3$ if postponement of flowering leads to a much higher seed production. Under natural conditions, such life history parameters will evolve to optimal combinations, given the trade offs. So, we would expect a low probability of flowering in the first year when seed production at an older age is relatively high and vice versa, leading to high hazard rates. Cultivar genes might further increase it by lowering the first year flowering probability of hybrids.

As long as adult plant survival chances are positive, their magnitude does not make much difference for the hazard rate, and the same holds for the flowering probability of two-year and older plants, $r_4$ (Figs. 5a,b,d). The effects of expected seed numbers are much more pronounced (Figs. 5e,f). In addition, the shapes of seed production distributions are also important, for instance, a higher variance decreases the hazard rate (22). As illustrated in Fig. 6, this effect can be significant. We expect this result to hold generally, regardless of the details of the model. Changes in variance of backcross seed production have larger consequences than those in variance of hybrid seed production (e.g., Fig. 6). This result is intuitively clear, since all subsequent generations of offspring from backcrossed individuals are governed by the backcrossed life-history parameters, whereas hybrid life-history parameters only affect one generation.

Based on these results we would advise that in the context of introgression risk, it is better to study effects of hybridization and backcrossing on seed production distributions and first-year flowering probability rather than probabilities of adult survival and flowering at older ages. Since we considered a very simplified model, however, these conclusions are only tentative.

We expect that an important application of the hazard rate will be the study of effects of time-varying environmental conditions, or crop management. In such cases there can be periods where the hazard rate decreases, for instance when crops are rotated. Methods for calculating hazard rates from time-inhomogeneous branching processes (Smith and Wilkinson, 1969) are currently being developed.

The model can be extended easily to incorporate other types of life histories and other modes of density dependent competition, provided that the wild population is large and homogeneous. As long as this holds, interactions between individuals with introgressed genotypes can be ignored initially and branching process models can be used to study their invasion dynamics. In small or spatially structured populations, invaders may affect each other already at low numbers. This necessitates the use of frequency- and density-dependent invasion models. Until now there are not many mathematical results on such generalizations of branching processes (but see Jagers and Klebaner, 2000). Furthermore, foreign genes invading small wild populations may become established by drift, even without a fitness advantage (for $q = 1$), especially when there are repeated invasions (Haygood et al., 2003). Methods for quantifying introgression processes in such situations remain to be developed.

Even in large wild populations, if invasion is successful, the density of invaders will eventually become so large that the possibility that invaders interact directly cannot be neglected. The invasion model that we used can be considered as an approximation of a more complicated model that includes such interactions, valid at low invader densities. In this light, $1?\,q$ should be considered as an approximation of the probability that the numbers of type-5 and 6 individuals reach such high levels that the probability that the foreign gene disappears from the population due to demographic stochasticity can be neglected. This type of approach is common in invasion studies (e.g., Garnier and Lecomte, 2006), and generally works well (see Champagnat et al., 2006).

We did not incorporate explicit genetics into the model, but obviously this is an important generalization, which is, for instance, needed to study effects of linkage and hitchhiking. A huge number of types may be needed to represent the different possible introgressed genotypes. Models can be considerably simplified, however, if some genotypes have equal fitnesses. In our model, for instance, we assumed that individuals of BC1 and later backcross generations have the same fitness.

Important other issues for future research are the effects of spatial structure and gene flow between subpopulations (e.g., Hanski, 1999) on the hazard rate.

In conclusion, further development of stochastic models for introgression research is needed, and we expect that this will require much more work. However, it is imperative to use such models, because, as we showed, deterministic models ignore important factors, and give misleading results. To analyze and interpret results of stochastic models, and go beyond simulations, we need measures such as the hazard rate, which adequately quantify essential features of stochastic introgression processes.

## Appendix A. Appendix

A.1. **Derivation of (15).** Here, we will use a multi-dimensional extension of the definition of the p.g.f. given in the main text, to deal with multiple types. Thus, the p.g.f. of the offspring of a single type-$i$ individual ($i \in \{0, 3, 4, 5\}$) is defined as:

$$F_i\left(s_0, s_3, s_4, s_5\right) = E\left[s_0^{Z_0(1)} s_3^{Z_3(1)} s_4^{Z_4(1)} s_5^{Z_5(1)} \left| Z_i\left(0\right) = 1, Z_j\left(0\right) = 0 \text{ for } j \neq i\right.\right] \quad \text{(A.1)}$$

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

$$F_0\left(s_0, s_3, s_4, s_5\right) = s_0 G_0\left(\hat{p}_0 s_3 + (1-\hat{p}_0)\right) \tag{A.2}$$

since a type-0 individual produces one of its own type, and a random number of seeds according to a p.g.f. $G_0(s)$, of which a proportion $\hat{p}_0$ of the seeds will germinate.

$$F_3\left(s_0, s_3, s_4, s_5\right) = r_3 G_3\left(\hat{p}_0 s_5 + (1-\hat{p}_0)\right) + (1-r_3)\, p_3 s_4 + (1-r_3)\,(1-p_3) \tag{A.3}$$

since a type-3 individual may flower with a probability $r_3$ and produce some number of seeds according to a p.g.f. $G_3(s)$, of which a proportion $\hat{p}_0$ will flower to become type-5 individuals.

Following a similar reasoning, we also obtain the following expression:

$$F_4\left(s_0, s_3, s_4, s_5\right) = r_4 G_4\left(\hat{p}_0 s_5 + (1-\hat{p}_0)\right) + (1-r_4)\, p_4 s_4 + (1-r_4)\,(1-p_4) \tag{A.4}$$

Next, we consider the generating function of $I_i(n)$:

$$\begin{aligned} f_{I_i(n)}(s) &= E\left[s^{I_i(n)}\right] \\ &= E\left[E\left[s^{I_i(n)} \mid Z_0(1), Z_3(1), Z_4(1), Z_5(1)\right]\right] \\ &= E\left[E\left[s^{\sum_{k=1}^{Z_0(1)} I_0(n-1)^{(k)} \sum_{k=1}^{Z_3(1)} I_3(n-1)^{(k)} \sum_{k=1}^{Z_4(1)} I_4(n-1)^{(k)} + Z_5(1)} \right.\right. \\ &\qquad \left.\left. \Big| Z_0(1), \ldots, Z_5(1)\right] \Big| Z_i(0) = 1, Z_j(0) = 0 \text{ for } j \neq i\right] \end{aligned} \tag{A.5}$$

where $I_j(n-1)(k)$ denotes the total number of type-5 individuals up to and including the next $n-1$ generations produced by type-3 and 4 individuals of the lineage of the $k$th individual of type $j$ in generation 1. Using the fact that individuals reproduce independently, and that individuals of the same type have identical offspring distributions, we can rewrite (A.5) as follows:

$$\begin{aligned} &E\left[E\left[s^{I_0(n-1)}\right]^{Z_0(1)} E\left[s^{I_3(n-1)}\right]^{Z_3(1)} E\left[s^{I_4(n-1)}\right]^{Z_4(1)} s^{Z_5(1)} \right. \\ &\qquad \left. \Big| Z_i(0) = 1, Z_j(0) = 0 \text{ for } j \neq i\right] \end{aligned} \tag{A.6}$$

and this equals

$$\begin{aligned} &E\left[\left(f_{I_0(n-1)}(s)\right)^{Z_0(1)} \left(f_{I_3(n-1)}(s)\right)^{Z_3(1)} \left(f_{I_4(n-1)}(s)\right)^{Z_4(1)} s^{Z_5(1)} \Big| \right. \\ &\qquad \left. Z_i(0) = 1, Z_j(0) = 0 \text{ for } j \neq i\right] \\ &\quad = F_i\left(f_{I_0(n-1)}(s), f_{I_3(n-1)}(s), f_{I_4(n-1)}(s), s\right) \end{aligned} \tag{A.7}$$

where we have used (A.1) in the last line. Using the above relationships, the equations of (15) follow.

A.2. **Derivation of Eq. (17).** Note that the last equality in Eq. (15) can be written as:

$$f_{I_4(n)}(q) = a_4(q) + b_4 f_{I_4(n-1)}(q) \tag{A.8}$$

with $a_4(q)$ and $b_4$ as defined in Eq. (11). Solving this recursion yields:

$$\begin{aligned} f_{I_4(n)}(q) &= a_4(q) + b_4 f_{I_4(n-1)}(q) = a_4(q) + a_4(q)\, b_4 + {b_4}^2 f_{I_4(n-2)}(q) \\ &= a_4(q) + a_4(q)\, b_4 + a_4(q)\, {b_4}^2 + {b_4}^3 f_{I_4(n-3)}(q) = \ldots = {b_4}^n + \sum_{j=0}^{n-1} a_4(q)\, {b_4}^j \\ &= a_4(q) \left( \frac{1 - {b_4}^n}{1 - b_4} \right) + {b_4}^n \end{aligned} \tag{A.9}$$

Note that if we use $n = 0$ we get $f_{I_4(0)}(q) = 1$, so the expression derived in (A.9) is valid for all $n$. Combining (A.9) with (15) and (16) gives Eq. (17).

A.3. **Effects of $r_4$, $p_3$ and $p_4$ on the asymptotic hazard rate.** The argument of $G_0$ in the expression of the asymptotic hazard rate in (18) is:

$$\hat{p}_0 \left( \begin{array}{c} r_3 G_3(\hat{p}_0 q + (1 - \hat{p}_0)) + (1 - r_3)(1 - p_3) \\ + \frac{(1-r_3)p_3}{1-(1-r_4)p_4} \left( r_r G_4(\hat{p}_0 q + (1 - \hat{p}_0)) + (1 - r_4)(1 - p_4) \right) \end{array} \right) + (1 - \hat{p}_0) \tag{A.10}$$

Furthermore, $G_0(s)$ is a monotonically increasing function of $s$. Therefore, if the expression in (A.10) increases with a certain parameter, then the asymptotic hazard rate decreases, and vice versa. The derivative of (A.10) with respect to $p_4$ is positive if

$$\begin{aligned} &-(1 - (1 - r_4)\, p_4) + (r_4 G_4(\hat{p}_0 q + (1 - \hat{p}_0)) + (1 - r_4)(1 - p_4)) > 0 \Leftrightarrow \\ &r_4 G_4(\hat{p}_0 q + (1 - \hat{p}_0)) - r_4 > 0 \Leftrightarrow G_4(\hat{p}_0 q + (1 - \hat{p}_0)) > 1 \end{aligned} \tag{A.11}$$

which gives a contradiction. We can conclude that (A.10) is monotonically non-increasing with $p_4$, and thus the asymptotic hazard rate must be monotonically non-decreasing in $p_4$. Results for $p_3$ and $r_4$ can be derived in a similar way.

A.4. **Derivation of (22).** Consider $G(s)$ to represent the p.g.f. of some random variable $X$. It can easily be shown that:

$$G(1) = 1, G'(1) = \mathrm{E}[X], \; G''(1) = \mathrm{E}[X(X-1)] \tag{A.12}$$

Looking at (18) and its derivatives at $q = 1$, we find the following expressions.

$$\hat{H}(1) = 0 \tag{A.13}$$

$$\hat{H}'(1) = -\hat{p}_0^2 m_0 \left( r_3 m_3 + \frac{b_3}{1 - b_4} r_4 m_4 \right) \tag{A.14}$$

$$\hat{H}''(1) = -\beta_0 \hat{p}_0^4 \left( r_3 m_3 + \frac{b_3}{1 - b_4} r_4 m_4 \right)^2 - m_0 \hat{p}_0^3 \left( r_3 \beta_3 + \frac{b_3 r_4}{1 - b_4} \beta_4 \right) \tag{A.15}$$

where $\beta_i$, represents the second derivative of $G_i(s)$ evaluated at 1. Equation (22) follows from substituting these results in the second order Taylor expansion of $\hat{H}(q)$ around the point 1.

## References

Abbott, R.J., James, J.K., Milne, R.I., Gillies, A.C.M., 2003. Plant introductions, hybridization and gene ow. Phil. Trans. Roy. Soc. Lond. B Biol. Sci. 358, 11231132.

Allendorf, F.W., Leary, R.F., Spruell, P., Wenburg, J.K., 2001. The problems with hybrids: setting conservation guidelines. Trends Ecol. Evol. 16, 613-622.

Athreya, K.B., Ney, P.E., 1972. Branching Processes. Springer, Berlin.

Champagnat, N., Ferriere, R., Mlard, S., 2006. Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models. Theor. Popul. Biol. 69, 297321.

Davis, S.A., Catchpole, E.A., Pech, R.P., 1999. Models for the introgression of a transgene into a wild population within a stochastic environment, with applications to pest control. Ecol. Model. 119, 267275.

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.

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, 539-63.

Ellstrand, N.C., 2003. Dangerous Liaisons? When cultivated plants mate with their wild relatives. The John Hopkins University Press, Baltimore.

Eshel, I., 1981. On the survival probability of a slightly advantageous mutant gene with a general distribution of progeny size: A branching process model. J. Math. Biol. 12, 355-362.

Feller, W., 1968. An Introduction to Probability Theory and its Applications. Third Edition, Volume 1. John Wiley and Sons, New York.

Garnier, A., Lecomte, J., 2006. Using a spatial and stage-structured invasion model to assess the spread of feral populations of transgenic oilseed rape. Ecol. Model. 194, 141149.

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

Haccou, P., Jagers, P., Vatutin, V.A., 2005. Branching Processes: Variation, Growth and Extinction of Populations. Cambridge University Press, Cambridge.

Hails, R.S., Morley, K., 2005. Genes invading new populations: a risk assessment perspective. Trends Ecol. Evol. 20, 245-252.

Haldane, J.B.S., 1927. A mathematical theory of natural and artificial selection. V. Selection and mutation. Proc. Cambridge Philos. Soc. 23, 838-844.

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

Hauser, T.P., Shaw, R.G., Ostergard, H., 1998. Fitness of hybrids between weedy Brassica rapa oilseed rape (B. napus). Heredity, 81, 436-443.

Haygood, R., Ives, A.R., Andow, D.A., 2003. Consequences of recurrent gene flow from crops to wild relatives. Proc. Biol. Sci. 270, 1879-1896.

Haygood, R., Ives, A.R., Andow, D.A., 2004. Population genetics of transgene containment. Ecol. Lett. 7, 213-220.

Huxel, G. R., 1999. Rapid displacement of native species by invasive species: effects of hybridization. Biol. Conservat. 89, 143-152.

Jagers, P., Klebaner, F.C., 2000. Population-size-dependent and age-dependent branching processes. Stochastic Process. Appl. 87, 35254.

Jenczewski, E., Ronfort, J., Chvre, A., 2003. Crop-to-wil gene flow, introgression and possible fitness effects of transgenes. Environ. Biosafety Res. 2, 9-24.

Kalbfleisch, J.D., Prentice, R.L., 2002. The statistical analysis of failure time data. 2nd edition. John Wiley & Sons, New York.

Kelly, C.K., Bowler, M.J., Breden, F., Fenner, M., Poppy, G.M., 2005. An analytical model assessing the potential threat to natural habitats from insect resistance transgenes. Proc. Biol. Sci. 272, 1759-1767.

Levin, D.A., Francisco-Ortega, J., Jansen, R.K., 1996. Hybridizationand the extinction of rare plant species. Conserv. Biol. 10, 1016.

Maan, S.S., 1987. Interspecific and intergeneric hybridisation in wheat. In Heyne, E.G., (Ed.), Wheat and wheat improvement. ASA, CSSA and SSSA, Madison, 453-461.

Michor, F., Nowak, M.A., Iwasa, Y., 2006. Stochastic dynamics of metastasis formation. J. Theor. Biol. 240, 521-530.

Reichman, J.R., Watrud, L.S., Lee, E.H., Burdick, C.A., et al., 2006. Establishment of transgenic herbicide-resistant creeping bentgrass (Agrostis stolonifera L.) in nonagronomic habitats. Mol. Ecol. 15, 13, 4243-4255.

Rieger, M.A., Lamond, M., Preston, C., Powles, S.B., Roush, R.T., 2002. Pollen-mediated movement of herbicide resistance between commercial canola fields. Science, 296, 2386-2388.

Riesberg, L.H., Wendel, J.F., 1993. Introgression and its consequences in plants. In Harrison, R.G., (Ed.), Hybrid Zones and the Evolutionary Process. Oxford University Press, Oxford, pp 70-109.

Serra, M.C., 2006. On the waiting time to escape. J. Appl. Prob. 43, 296-302

Serra, M.C., Haccou, P., 2007. Dynamics of escape mutants. Theor. Popul. Biol., 72, 167-178.

Smith, W.L., Wilkinson, W.E., 1969. On branching processes in random environments. Ann. Math. Statist. 40, 814827.

Snow, A.A., Andersen, B., Jorgensen, R.B., 1999. Costs of transgenic herbicide resistance introgressed from Brassica napus into weedy B. rapa. Mol. Ecol. 8, 605-15.

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.

# CHAPTER 2: COMBINING MODELS WITH EXPERIMENTS

## 1. Introduction

The models presented within this thesis have been developed as a part of a wider project which uses the carrot (*Daucus carota*) as a case study for the development of a methodology to quantify introgression risk realistically. The carrot is primarily an outcrossing species, and there have been many documented occurrences of hybridisation between wild and cultivated carrots (Wijnheijmer et al. 1989 Hauser and Bjorn 2001, for example), so it is a good candidate for such a study. The development of models has been concurrent with the execution of field trials by collaborators working on the same project. At the time of writing, field trials are ongoing, but have reached a stage where the models in the thesis can be combined with preliminary results from experiments.

The data for the calculations come from a field trial conducted during the winter of 2011 in Lisse, The Netherlands. A commercially available variety known as Flakkese was used as a cultivar, which was crossed and backcrossed with wild carrots collected from Stevenshof, a suburb of Leiden, The Netherlands. For full details of crossing experiments, and field trials, see Grebenstein (in prep.). A summary of this data for wild plants, F1 hybrids, and BC1 backcrossed individuals can be found in Table 1. The data set at the time of writing is still incomplete, with data for umbel sizes available as opposed to full seed sets, small sample sizes, and some missing data. I use them merely as an illustration of how the hazard rate is calculated from real data.

Table 1. Summary of data used

| Plant type | Average primary umbel diameter (cm) | Average primary umbel area ($cm^2$) | Survival probability to flowering |
|---|---|---|---|
| Stevenshof | 5.73 | 27.52 | 0.94 |
| F1 | 6.72 | 40.70 | 1 |
| BC1 | 6.25 | 33.61 | 0.86 |

The average umbel area was calculated from the separately measured umbel diameters, under the assumption that each umbel is circular.

## 2. Calculation of the hazard rate

All plants in the field trial either flowered or died after one year, i.e. no flowers remained in a vegetative state. Thus, we use the model in Chapter 1 with the survival probabilities of non-flowering one year old plants set to zero (i.e. $p_1 = p_3 = p_5 = 0$). Please refer to Chapter 1 for full derivations and details of assumptions.

To estimate the seed set from the data on umbel diameter, first umbel area was calculated, assuming that each umbel was circular. From the umbel area, a

measure for the seed set was calculated, assuming that each plant had the same density of seeds per unit of umbel area, denoted by the constant, $k$.

2.1. **Seed establishment probability.** Adapting previous results to annuals leads to the following seed establishment probability of a single seed:

$$\hat{p}_0 = \frac{1}{m_1 r_1}, \tag{1}$$

where $r_1$ is the flowering probability of a wild plant, and $m_1$ is the expected number of seeds that a flowering wild plant produces. Looking at the values in Table 1 and substituting into Eq. (1), we find $\hat{p}_0 = \frac{1}{27.52*0.94*k} = \frac{1}{25.87k}$.

2.2. **Extinction probability of a lineage initiated by a backcrossed individual.** The extinction probability of the lineage initiated by a single backcrossed becomes the following when made applicable for annuals:

$$q = r_5 G_5\left(\hat{p}_0 q + (1 - \hat{p}_0)\right) + 1 - r_5. \tag{2}$$

where $r_5$ is the flowering probability of a single backcrossed individual. $G_5(s)$ is the probability generating function (p.g.f.) of the seed numbers produced by a single BC1 plant. Using a Poisson p.g.f. (i.e. $G_5(s) = e^{-m_5(1-s)}$, where $m_5$ is the expected number of seeds produced by a flowering BC1 plant) yields the following expression:

$$q = r_5 e^{-\hat{p}_0 m_5 (1-q)} + 1 - r_5. \tag{3}$$

From Table 1, we have $m_5 = 33.61k$. Putting this value into the above equation, and also substituting the calculated value of $\hat{p}_0$, the following expression is reached.:

$$q = 0.86 e^{-\frac{33.61}{25.87}(1-q)} + 0.14. \tag{4}$$

Note that this expression is independent of $k$. Also, the solution of $q = 1$ satisfies this equation, but the extinction probability is the smallest root of this equation, which can be calculated numerically to give $q = 0.83$ (to two decimal places).

2.3. **The asymptotic hazard rate.** In this case the asymptotic hazard rate equals

$$\hat{H}(q) = 1 - G_0\left(\hat{p}_0\left(r_3 G_3\left(\hat{p}_0 q + 1 - \hat{p}_0\right) + 1 - r_3\right) + 1 - \hat{p}_0\right) \tag{5}$$

where $G_0(s)$ is the p.g.f. of the number of hybrid seeds produced in the wild population per generation, $G_3(s)$ is the p.g.f. of the number of seeds produced by a flowering F1 hybrid, and $r_3$ is the flowering probability of an F1 hybrid. It is easiest to break up the calculation of Eq. (5) into two parts. First, define and calculate a part of the argument of $G_0(s)$ as follows:

$$\begin{aligned} c &= r_3 G_3\left(\hat{p}_0 q + 1 - \hat{p}_0\right) + 1 - r_3 \\ &= r_3 e^{-\hat{p}_0 m_3 (1-q)} + 1 - r_3 \end{aligned} \tag{6}$$

where I have assumed that the seed production of a flowering F1 plant is Poisson distributed in writing down the second line. From Table 1, we have that $r_3 = 1$ and $m_3 = 40.70k$. This gives a value of $c = 0.76$ (to two decimal places). Note that $c$ is independent of $k$.

Assuming that the number of hybrid seeds produced in the wild population is Poisson-distributed with mean $m_0$, we find the following expression for the asymptotic hazard rate:

$$\begin{aligned}\hat{H}(q) &= 1 - e^{-\hat{p}_0 m_0 (1-c)} \\ &= 1 - e^{-m_{hyb}(1-c)}\end{aligned} \tag{7}$$

The term $\hat{p}_0 m_0$ in the exponent is the product of the seed-establishment probability and the expected number of hybrid seeds formed per generation. Thus, $\hat{p}_0 m_0$ can be interpreted as the expected number of hybrid plants produced per generation, denoted by $m_{hyb}$. This can vary, depending on several factors, e.g. distance the wild population to the crop field, and size of the wild population.

## 3. Results

Figure 1 shows the asymptotic hazard rate plotted against hybridisation rate.At small hybridisation rates, as in Fig. 1 (b), the asymptotic hazard rate increases nearly linearly, and we can use the first order Taylor approximation:

$$H \approx m_{hyb}(1-c). \tag{8}$$

At larger hybridisation rates, the hazard rate approaches one. This can be seen straight from Eq. (7), where the exponential term becomes zero as $m_{hyb}$ becomes large.

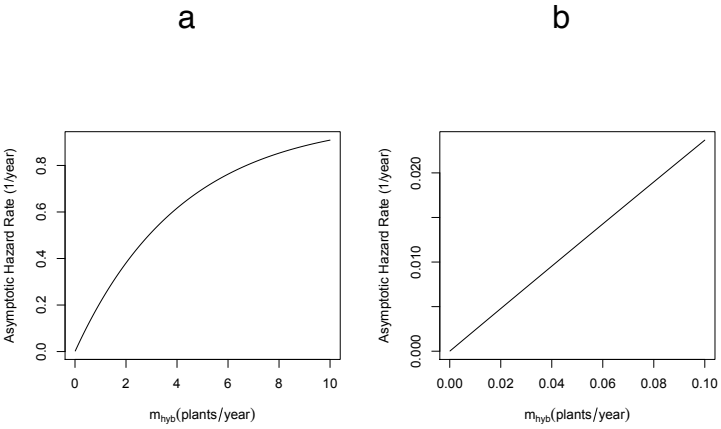


Figure 1. The asymptotic hazard rate plotted against $m_{hyb}$ at two different scales. Parameters are as shown in Table 1.

Figure 2 shows the effect that changing the average F1 and BC1 umbel areas has on the asymptotic hazard rate. The asymptotic hazard rate is more sensitive to changes in BC1 fitness than F1 fitness.

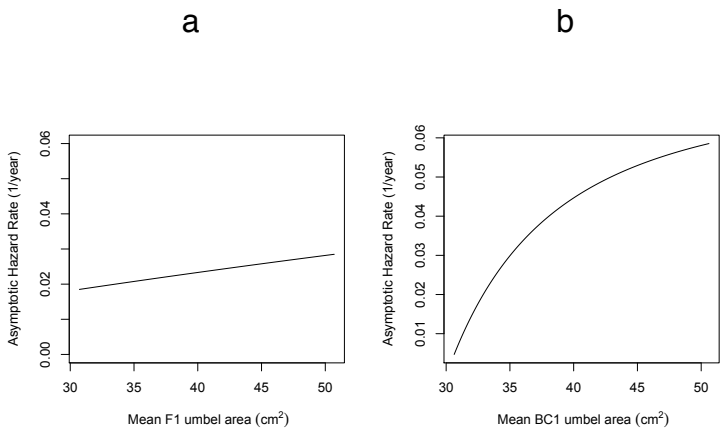


FIGURE 2. The asymptotic hazard rate plotted agains average F1 (subplot (a)) and BC1 (subplot (b)) primary umbel areas. A value of $m_{hyb} = 0.1$ is used for both plots. Other relevant parameters are as shown in Table 1.

## 4. Discussion

In this chapter, I have applied the methodology from Chapter 1 to estimate the hazard rate of introgression of crop into genes using measured data from a field trial of *Daucus carota*. Data on hybridisation rates is still required to accurately quantify introgression risk, and this data is currently being gathered. On the basis of the calculations presented here, we can predict that an average of 0.1 hybrid plants produced per generation in the wild population will lead to a hazard rate of 0.024. This value implies that the expected time until introgression (initiation of a permanent lineage) occurs is 40 years.

The sensitivity analysis shows that the asymptotic hazard rate is more sensitive to BC1 fitness than F1 fitness. The reason for this can be explained by an assumption made in the Chapter 1, namely that BC2 and subsequent backcross generations have the same life-history parameters as BC1 plants. Consequently, the life-history parameters of BC1 plants affects all subsequent generations and the model is especially sensitive to changes in BC1 fitness. If data about BC2 and further backcross generations is gathered empirically, then this can be incorporated into the model, and procedures for doing so are given in Chapter 3.

To calculate the hazard rate from the currently avaliable data, several assumptions had to be made, in addition to the assumptions already enumerated in Chapter 1. The data is only currently available in the form primary umbel diameters, whereas data is required in the form of seed sets. Consequently, a direct proportionality was assumed between umbel area and seed set. Furthermore, I assumed the same constant of proportionality to hold for wild, F1 and BC1 plants. This might not be the case, e.g. area in larger umbels might be due to empty space and not seeds. I also assumed that the numbers of seeds produced by plants are

Poisson-distributed. This is a convenient choice, since it results in the final result of the hazard rate to be independent of the constant of proportionality, $k$. An alternative would be to use p.g.f.s as implied by umbel areas, but this would allow an arbitrary choice of $k$ in the final result, so would be unsatisfactory in that regard.

In a full implementation of the model, $\hat{p}_0$ would be calculated from the life-history parameters of wild plants, and $m_0$ would be measured. A hurdle in the implementation in this chapter is that there currently no measurement for $m_0$, and $\hat{p}_0$ is dependent on $k$. It was possible to combine these two parameters and interpret the combination as the expected number of hybrid plants produced per generation. Results could be seen in terms of this hybridisation rate. In a full implementation of the model, using seed sets instead of umbel area would result in no factor of $k$ appearing in $\hat{p}_0$, and data would be available for $m_0$, so these assumptions could be avoided. With richer and more complete data sets, more accurate estimates for introgression risk can be made. Work towards this is currently underway.

## References

Grebenstein, C., in prep.

Hauser, T.P., Bjorn, G.K., 2001. Hybrids between wild and cultivated carrots in Danish carrot fields. Genet. Resour. Crop Ev., 48, 499-506.

Wijnheijmer E.H.M., Brandenburg, W.A., Terborg, S.J., 1989. Interactions between wild and cultivated carrot (*Daucus carota* L) in the Netherlands. Euphytica, 40, 147-154.

# CHAPTER 3: QUANTIFYING TIME-INHOMOGENEOUS STOCHASTIC INTROGRESSION PROCESSES WITH HAZARD RATES

Reprinted with minor edits from Ghosh, Haccou, 2012. Theor. Popul. Biol., 81, 253-263

## Abstract

Introgression is the permanent incorporation of genes from one population into another through hybridization and backcrossing. It is currently of particular concern as a possible mechanism for the spread of modified crop genes to wild populations. The hazard rate is the probability per time unit that such an escape takes place, given that it has not happened before. It is a quantitative measure of introgression risk that takes the stochastic elements inherent in introgression processes into account. We present a methodology to calculate the hazard rate for situations with time-varying gene flow from a crop to a large recipient wild population. As an illustration, several types of time-inhomogeneity are examined, including deterministic periodicity as well as random variation. Furthermore, we examine the effects of an extended fitness bottleneck of hybrids and backcrosses in combination with time-varying gene flow. It is found that bottlenecks decrease the hazard rate, but also slow down and delay its changes in reaction to changes in gene flow. Furthermore, we find that random variation in gene flow generates a lower hazard rate than analogous deterministic variation. We discuss the implications of our findings for crop management and introgression risk assessment.

## 1. Introduction

Through backcrossing and hybridization, genes from one population can become permanently incorporated into the genome of another population. This process is called introgression (Riesberg and Wendel, 1993; Ellstrand et al., 1999; Hails and Morley, 2005). Introgression of crop genes into wild relatives may have severe negative environmental effects, such as the spread of insecticide or herbicide resistance genes. In particular, there are strong concerns about transgene escape and its consequences, e.g. the production of superweeds (Maan, 1987; Snow et al., 1999; Thompson et al., 2003; Kelly et al., 2005).

The likelihood of such scenarios, given environmental conditions, crop management, and characteristics of the species involved can be studied with mathematical models. Such models allow us to perform thought experiments, and identify factors that crucially determine introgression risk. Introgression usually involves many random components, such as hybridization and backcross events, and demographic stohasticity in hybrid populations. In a previous paper (Ghosh and Haccou, 2010) we showed that it is important to take this stochasticity into account, since stochastic models may give very different predictions from deterministic ones. We considered a situation where foreign genes invade repeatedly

into a resident wild population, and each invasion has a small probability of establishing a permanent lineage (see also Haygood et al., 2004). We showed that there can be an extensive period of failed invasions, and that the length of this period largely determines introgression risk. Furthermore, we derived a measure, the hazard rate, that quantifies the distribution of such periods. In the context of introgression, the hazard rate is defined as the probability per time unit that a permanent lineage is initiated, given that this has not happened before. It is derived from a multitype branching process model of hybrid population dynamics (Demon et al., 2007; Serra and Haccou, 2007).

In our previous paper we assumed that the distribution of numbers of newly created hybrids is the same in each time period. We considered a model with an initial fitness bottleneck (i.e. F1 hybrids have a lower fitness than the wild type) and showed that in such a situation, the hazard rate increases monotonically from zero to a constant asymptotic value. As a consequence, the distribution of the initial period before establishment of a permanent lineage can be approximated by a time-lagged geometric distribution. In many applications, however, the hybridization probability will vary in time, due to, for example, crop rotation or termination, or random variation, such as weather-dependent pollinator activity. In the current paper we generalize the method to include such time-inhomogeneity. We calculate the hazard rate for general time-inhomogeneous hybridization schemes and examine the effects of crop management schemes such as (gradually) stopping or increasing crop cultivation, or rotating crops. We show that, in the latter case, periods in which the hazard rate increases alternate with periods of decrease, and that, in the long run, it converges to a periodic function. We also examine how stochastic fluctuations in hybridization rates affect the hazard rate.

As an example we consider a model for a monocarpic species (it dies after flowering), that is monoecious (flowers have both male and female functions), and non-selfing. We first consider a situation where F1 hybrids have a reduced fitness when compared to the wild-type, and all backcrosses have the same life history parameters, and superior fitness. Then the model is generalized to examine the effects of an extended fitness bottleneck, where several initial backcross generations have a reduced fitness.

There are many other contexts in which repeated invasions with low initial fitness occur, such as tumor spread and growth, where usually several mutations must occur before cells proliferate (as in Michor et al., 2006), or pathogen host switching, where adjustments to new hosts imply an initial fitness bottleneck (as in Reluga et al., 2007). Time-inhomogeneity of invasions may play a role in such contexts too. For instance, there may be time-varying risks of exposure to carcinogenic environments (e.g. Bos et al., 2004). Furthermore, many epidemics show time-varying infection patterns (as in Welliver, 2009). Our methods and results therefore have implications for research in such contexs too.

## 2. The model

We consider a plant species that dies after flowering once. For simplicity, we assume that there is no age-dependence. Furthermore, it is assumed that there is a large, stable wild population, and random numbers of hybrid seeds are produced by pollen flow from a nearby crop. We consider time periods of one year. Seeds

may germinate at the beginning of the year, and plants grow up to be adults and may flower later in the same year. We denote the probability that a seed germinates and that the seedling survives to become an adult plant by $p_0$. In this paper we will consider $p_0$ as a given parameter. Its value is determined by the population dynamics of the wild population, and is such that this population is stable (see Ghosh and Haccou, 2010, for an example of its calculation).

Hybrid formation can be followed by repeated backcrossing with wild plants. F1 hybrids are assumed to be less fit than wild individuals, but backcrossed individuals have a positive probability of producing a permanent introgressed lineage. We assume that all backcross generations are equivalent with respect to their life history parameters, and therefore they do not need to be distinguished as seperate types (this assumption is relaxed in section 6). As a consequence, there are two types of plants in the model: F1 hybrids (labelled type-1) and backcrossed individuals (labelled type-$E$).

Since the population of wild plants is large and the numbers of individuals containing crop genes are initially small, it can be assumed that these individuals do not interact with each other, but only with wild plants. This has several implications. Firstly, since we consider a non-selfing species, reproduction can only occur through outcrossing with wild plants. Secondly, competition occurs only with the wild population. This is quantified through the probability $p_0$. For convenience, we assume that there are no other factors apart from this competition that affect germination probability of hybrids and backcrosses. The model can be easily generalized to account for e.g. effects of spatial variation.

Because hybrid and backcrossed plants do not affect each other's reproduction and survival initially, their invasion dynamics can be modeled as a branching process. The production of hybrid seeds is modeled by means of an artificial type, which we will call type-0. There is one permanently present individual of this type, that produces a stochastic number of hybrid seeds in each year. Fig. 1 shows a schematic summary of the invasion dynamics.

The model thus involves three different types of individuals: type-0, type-1 and type-$E$. Each year, a type-0 individual produces one individual of type-0 and a random number of F1 hybrid seeds. In our previous paper we assumed that the probability distribution of these random numbers was the same over time. In this paper, we let it vary over years. The number of hybrid seeds produced in year $k$ is a random variable denoted by $\xi_{0,k}$. Each one of these seeds germinates and produces a type-1 individual with probability $p_0$. Type-1 individuals flower with probability $r_1$, and produce a random number, $\xi_1$, of backcrossed seeds, either by male or female functions. In the case that a type-1 individual does not flower (with a probability $(1 - r_1)$), it may then survive to become a type-1 individual in the next year with probability $p_1$, or it will die with a probability $1 - p_1$. Each backcrossed seed germinates and survives with probability $p_0$, to produce a type-$E$ individual. Type-$E$ individuals produce only type-$E$ offspring in their lineage. We denote the probability that a lineage started by one type-$E$ individual goes extinct by $q$. This value can be calculated straightforwardly from the life history parameters of type-$E$ individuals, by standard methods (see e.g. Haccou et al., 2005; Ghosh and Haccou, 2010). Here, we will treat it as a parameter in the model, taking values between zero and one.

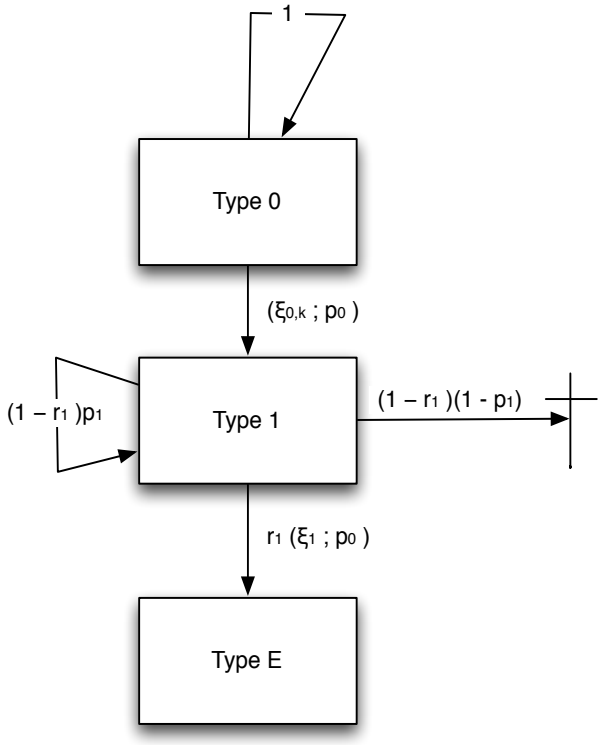


FIGURE 1. Schematic representation of the model. $(\xi_{0,k}; p_0)$ and $(\xi_1; p_0)$ represent the production of $\xi_{0,k}$ and $\xi_1$ seeds respectively, where each seed has a germination probability $p_0$. Each type-$E$ individual initiates a lineage which eventually becomes extinct with probability $q$.

## 3. Derivation of the hazard rate

Probability generating functions are important tools in deriving the hazard rate. Let $X$ be a non-negative discrete random variable, then its probability generating function (p.g.f.) is a function from $[0, 1]$ to $[0, 1]$ which is defined as $E[s^X]$, where $E[\,.\,]$ denotes expectation. The p.g.f. of $\xi_{0,k}$ is denoted by $G_0(k; s)$, and that of $\xi_1$ by $G_1(s)$.

Define the random variable $I_i\,(k, n)$ $(n, k \in \mathbb{N}_0,\ i = 0, 1)$ to be the total number of type-$E$ individuals with non type-$E$ parents, appearing up to and including year $n$, in the *line of descent* of a single individual of type-$i$ that was produced in year $k$. The expression *line of descent* refers to the population process stemming from the referred individual. For a general scenario where individuals can have

offspring of any type, this definition leads to the following equalities:

$$I_i(k,n) = \begin{cases} 0 & \text{if } k \geq n \\ Z_E^{(i)}(k+1) + \sum_{m=0}^{1} \sum_{j=1}^{Z_m^{(i)}(k+1)} I_m^{(j)}(k+1,n) & \text{if } k < n \end{cases} \quad (1)$$

where $Z_m^{(i)}(k+1)$ represents the number of type-$m$ individuals that the type-$i$ individual (born in year $k$) produced in year $k+1$. The $I_m^{(j)}(k+1,n)$ terms represent the total number of type-$E$ individuals that have non type-$E$ parents, appearing up to year $n$ in the line of descent of the $j^{th}$ individual of type-$m$ that was born in year $k+1$ from the initial type-$i$ individual.

In the specific scenario described in Fig. 1, we find the following recursive relationships in $k$ for the different p.g.f.'s of the $I_i(k,n)$'s, where $f_{I_i(k,n)}(s)$ denotes the p.g.f. of $I_i(k,n)$ (see Appendix A.1):

$$\begin{aligned} f_{I_0(k,n)}(s) &= f_{I_0(k+1,n)}(s) G_0\left(k; p_0 f_{I_1(k+1,n)}(s) + 1 - p_0\right) \\ f_{I_1(k,n)}(s) &= (1-r_1)(1-p_1) + (1-r_1)p_1 f_{I_1(k+1,n)}(s) + r_1 G_1(p_0 s + 1 - p_0) \end{aligned} \quad (2)$$

with the initial conditions $f_{I_1(n,n)}(s) = f_{I_0(n,n)}(s) = 1$. Note that, since the seed production of type-1 individuals is homogeneous,

$$f_{I_1(k,n)}(s) = f_{I_1(0,n-k)}(s). \quad (3)$$

The time of an introgression event, $T$, is defined as the time that the first type-$E$ individual is produced whose lineage never becomes extinct. The population starts with a single type-0 individual, therefore:

$$P(T > n) = f_{I_0(0,n)}(q), \quad (4)$$

since the probability that an introgression event occurs after a time $n$ is the probability that all type-$E$ individuals produced at or before year $n$ have become extinct.

The hazard rate of introgression is defined as the probability per time unit that an introgression event occurs given that it has not occurred before. With time units of one year, this gives:

$$H_n(q) = P(T = n | T > n-1)/year = \left(1 - \frac{f_{I_0(0,n)}(q)}{f_{I_0(0,n-1)}(q)}\right) year^{-1} \quad (5)$$

with $n \in \mathbb{N}_0$.

The second equation of (2) can be solved to yield (see Appendix A.2):

$$f_{I_1(0,n)}(s) = 1 - \beta_1(s) + \beta_1(s)\, b_1^n, \quad (6)$$

where, in order to simplify future expressions, we have introduced the quantities

$$b_1 = (1-r_1)\, p_1 \quad \text{and} \quad \beta_1(s) = \frac{r_1\,(1 - G_1(p_0 s + 1 - p_0))}{1 - b_1} \quad (7)$$

Putting (2), (3), (4) and (5) together gives us the following expression for the hazard rate (see Appendix A.3):

$$H_n(q) = \begin{cases} 0 & \text{if} \quad n \in \{0,1\} \\ 1 - \dfrac{\prod_{j=1}^{n-1} G_0\left(j-1; p_0 f_{I_1(0,n-j)}(q) + 1 - p_0\right)}{\prod_{j=1}^{n-2} G_0\left(j-1; p_0 f_{I_1(0,n-1-j)}(q) + 1 - p_0\right)} & \text{if} \quad n \geq 2 \end{cases} \tag{8}$$

which can be computed by using (6). This result provides us with a general method for calculating the hazard rate with time-inhomogeneous hybridization. In the next sections we examine several situations.

## 4. Deterministically varying hybridization

For mathematical convenience we assume that hybrids are generated according to a Poisson distribution with a time-dependent mean, i.e.:

$$G_0(k; s) = e^{-m_0(k)(1-s)},\ s \in [0,1]. \tag{9}$$

We also take $\xi_1$ as Poisson-distributed with mean $m_1$ in presented numerical work.

Combining (6) to (9) gives:

$$H_n(q) = \begin{cases} 0 & \text{if} \quad n \in \{0,1\} \\ 1 - e^{-p_0\beta_1(q)(1-b_1)b_1^{n-2}\sum_{j=0}^{n-2} m_0(j)\, b_1^{-j}} & \text{if} \quad n \geq 2. \end{cases} \tag{10}$$

From (10) it follows that the long term behaviour of the hazard rate depends on the limit behaviour, as $k \to \infty$, of:

$$b_1^k \sum_{j=0}^{k} \frac{m_0(j)}{b_1^j}.$$

For example, if $m_0(j) = m_0^j$, the hazard rate converges to zero when $0 < m_0 < 1$ and it converges to one when $m_0 > 1$. If there is constant hybridization, i.e $m_0(j) = m_0$, the hazard rate tends to a constant value between zero and one (as was also derived in Ghosh and Haccou, 2010). It can easily be shown that, for the current model, this value equals

$$1 - \exp\{-p_0\beta_1(q)m_0\}. \tag{11}$$

In the next subsections we will examine the effects of specific frequently used crop-management schemes.

4.1. **Temporary crops.** Crop cultivation may be stopped for a variety of reasons. In the case of transgene crops, e.g., legislation may change, or termination of cultivation may be used as a management strategy to lower the chance of introgression. In this sub-section we examine the case where hybridization occurs at a constant rate, and is then stopped at a fixed time $S$, i.e.:

$$m_0(j) = \begin{cases} m_0 & \text{if} \quad 0 \leq j < S \\ 0 & \text{if} \quad j \geq S, \end{cases} \tag{12}$$

with $m_0 > 0$.

Substituting this into (10) gives:

$$H_n(q) = \begin{cases} 0 & \text{if} \quad n \in \{0,1\} \\ 1 - e^{-m_0\, p_0\, \beta_1(q)\, (1-b_1^{n-1})} & \text{if} \quad 2 \le n \le S+1 \\ 1 - e^{-m_0\, p_0\, \beta_1(q)\, b_1^{n-(S+1)} \left(1-b_1^{S}\right)} & \text{if} \quad n \ge S+2 \end{cases} . \tag{13}$$

Thus, the hazard rate increases monotonically to a maximum level of $1 - e^{-m_0\, p_0\, \beta_1(q)\, (1-b_1^{S})}$ at time $S+1$ and decays monotonically afterwards. The decay is only seen to start at time $S+2$ because stopping hybridization at year $S$ will only affect the population of type-1 individuals at time $S+1$, and the population of type-$E$ individuals at time $S+2$. The rate of increase as well as that of decay is mainly governed by $b_1$, which represents the probability that individuals do not flower but do survive (see (7)). A larger value of $b_1$ makes the hazard rate increase and decrease more slowly. When $b_1$ tends to zero (i.e. when the probability of flowering in the first year is high and/or the survival probability of non-flowering adults is low), the maximum level is reached quickly and, unless $S$ is very small, it is therefore virtually independent of $S$. Furthermore, after stopping cultivation, the hazard rate returns rapidly to zero. As $b_1$ tends to zero or $S$ tends to infinity, the maximum level approaches the asymptotic level of the hazard rate in the situation without stopping. The effect of the life history parameters on this asymptotic level can be inferred from (11).

With temporary crops, there is a positive probability that introgression never occurs. From (4), (9), (12) and the derivation in Appendix A.3 it is apparent that this probability equals:

$$\lim_{n\to\infty} P(T > n) = \lim_{n\to\infty} f_{I_0(0,n)}(q) = e^{-m_0 p_0 \beta_1(q) S} \tag{14}$$

Thus, it decreases exponentially with the stopping time $S$, at a rate determined by the hybridization rate and the life history parameters.

A numerical example of the shape of the hazard rate for two different stopping times (10 and 20 years) is given in Fig. 2a. In this example, the hazard rate increases quickly, and, as a consequence, its maximum level does not noticeably differ for the two chosen stopping times. The probability distribution of $T$ can be expressed in terms of the hazard rate as follows (see e.g. Kalbfleisch and Prentice, 2002):

$$P\left(T = x\right) = \prod_{i=0}^{x-1} \left(1 - H_i(q)\right) H_x(q). \tag{15}$$

For small values of $H_n(q)$, the product term is close to one, and the probability becomes nearly equal to the hazard rate. This is demonstrated in Fig. 2b. As can be seen from the figure, the probabilities of introgression events happening quite early are relatively large, i.e. the probability distributions are very skewed, similar to the situation with constant crop cultivation examined before in Ghosh and Haccou (2010). For the numerical examples in Fig. 2b, the probabilities that no introgression occurs at all are respectively 0.985 ($S = 10$) and 0.970 ($S = 20$).

4.2. **Crop rotation.** Crop rotation is often used to maintain soil quality and prevent the build up of pathogens. It may also be used as a management strategy to lower introgression risk. In this section we study the situation where periods

with hybridization at a constant rate alternate with periods without hybridization. The duration of hybridization periods is denoted by $S$, and the durations of the *hybridization pauses* by $R$. Thus we have:

$$m_0(j) = \begin{cases} m_0 & \text{if} \quad v(R+S) \le j < v(R+S)+S \\ 0 & \text{if} \quad v(R+S)+S \le j < (v+1)(R+S) \end{cases} \tag{16}$$

with $v \in \mathbb{N}_0$.

It can be shown (see Appendix A.4) that in the long run the hazard rate tends to a periodic function with period $R+S$, i.e. if we define the time:

$$k = n - v(R+S) - 2 \tag{17}$$

then, for $n$ tends to infinity the hazard rate becomes:

$$\mathcal{H}_k(q) = \begin{cases} 1 - e^{-m_0 p_0 \beta_1(q)\left(1 - b_1^{k+1}\frac{1-b_1^R}{1-b_1^{R+S}}\right)} & \text{if} \quad 0 \le k < S \\ 1 - e^{-m_0 p_0 \beta_1(q) b_1^{k+1-S}\frac{\left(1-b_1^S\right)}{1-b_1^{(R+S)}}} & \text{if} \quad S \le k < R+S \end{cases} \tag{18}$$

The time in (17) is the time after the $v$th crop rotation shifted by two time units. The shift of two units is for mathematical convenience, and corresponds for the first two years where the hazard rate is zero.

This result implies that periods in which instantaneous introgression risk is high alternate with periods in which it is low. Figure 2c illustrates that this asymptotic behavior can be reached very quickly. Figure 2d shows the corresponding probabilities of introgression events happening at time $x$. As noted previously, the probability distribution is nearly equal to the hazard rate initially, but (inevitably) decreases with $x$.

There are different ways to quantify the effect of a given crop rotation scheme on the hazard rate. The asymptotic maximum hazard rate can be found by subsituting $k = S-1$ in (18), leading to:

$$1 - e^{-m_0 p_0 \beta_1(q)\frac{1-b_1^S}{1-b_1^{(R+S)}}}, \tag{19}$$

and the minimum by subsituting $k = R+S-1$, which gives:

$$1 - e^{-m_0 p_0 \beta_1(q) {b_1}^R \frac{1-b_1^S}{1-b_1^{(R+S)}}}. \tag{20}$$

For the numerical example in Figure 2c the asymptotic maximum hazard rate equals 0.00154, and the minimum is of the order $10^{-6}$. As can be seen from the figure, these values are reached quite soon.

An alternative measure is the long-run average hazard rate. This is found by fitting the survivor function of a constant hazard rate to the survivor function of the hazard rate from (18). This approach leads to the following value for the long-run average hazard rate (see Appendix A.5 for details):

$$\lambda \approx 1 - e^{-p_0 m_0 \beta_1(q)\frac{S}{R+S}}. \tag{21}$$

Thus, the long-run average hazard rate is the same as the asymptotic hazard rate with a continuous crop and a constant expected number of newly produced hybrids equal to $S/(R+S)$ times $m_0$. In Fig. 2d we have indicated the time-distributions

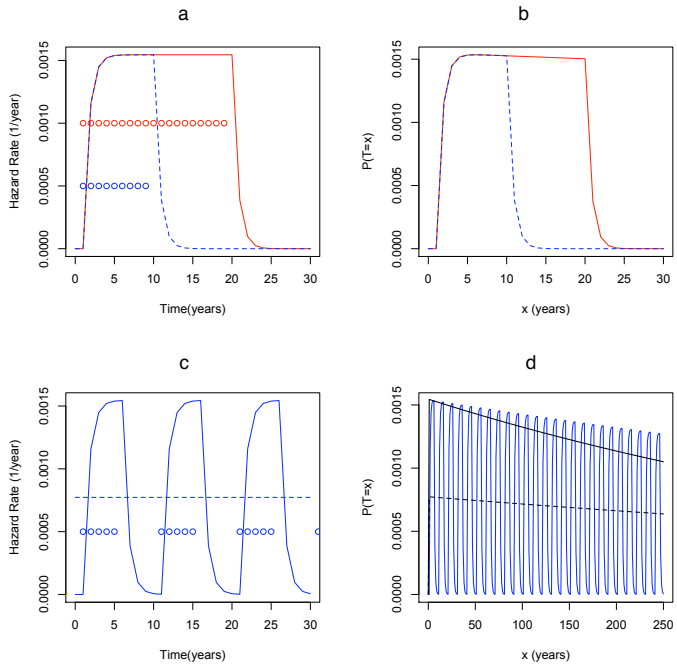


FIGURE 2. (a) Hazard rates when crops are terminated after a period of $S = 10$ (blue), or $S = 20$ (red). Parameter values: $m_0 = 50$, $p_0 = 0.001$, $p_1 = r_1 = 0.5$, $m_1 = 950$, $q = 0.95$, (b)The distributions of time until an introgression event, corresponding to the situations in (a). (c) The hazard rate with crop rotation (see (A.9) and (A.10)) (solid line) for $R = S = 5$ and all other relevant parameters the same as in (a). The average hazard rate (see (21)) (dotted line). (d) Distribution of times until an introgression event for the crop rotation scenario of (c) (blue line), for a constant average hazard rate (dotted black line), and for a constant maximum hazard rate (see (19), solid black line). In (a) and (c), circles indicate periods when hybridization occurs, but not the amount of immigration.

corresponding to a continuous immigration with the maximum hazard rate (c.f. (19)) and the long-run average hazard rate.

## 5. Randomly varying hybridization

Until now we have considered deterministic variation in hybridization rates. In many cases, however, there will also be random variation. For instance, weather conditions will vary over different years, and this may affect pollen dispersal from the crop to local wild populations. Such random variations can be independent, or (positively or negatively) autocorrelated. In this section, we consider the effect of random variation according to different regimes.

Random temporal variation of $m_0$ can be included in the model by using different type-0 individuals. Thus, we consider $\gamma$ different types, denoted by type-$(0,i)$ $(i = 1, ..., \gamma)$. A type-$(0,i)$ individual produces a number of type-1 seeds according to a p.g.f. $G_{0,i}(s)$, and with probability $\kappa_{i,j}$ also exactly one individual of type-$(0,j)$ $(j = 1, ..., \gamma)$, so $\sum_{j=1}^{\gamma} \kappa_{i,j} = 1$ for all $i$.

As an illustration, consider the case where the environment alternates between two states according to a two-type Markov chain. In that case $\gamma = 2$. When the environment is state 1, a Poisson-distributed number of hybrids is formed, i.e. $G_{0,1}(s) = e^{-m_0(1-s)}$ and when the environment is in state 2, no hybrids are produced, i.e. $G_{0,2}(s) = 1$. The transition probability from state 1 to state 2 equals $\kappa_{1,2}$ and that from state 2 to state 1 equals $\kappa_{2,1}$. An independently varying environment corresponds to the situation where $\kappa_{1,2} + \kappa_{2,1} = 1$. In the case of positive autocorrelation, this sum is smaller than one whereas it is larger than one for negatively autocorrelated environments.

As a special case, consider an independently varying environment, with $\kappa_{1,1} = \kappa_{2,1} = S/(R+S)$ and $\kappa_{1,2} = \kappa_{2,2} = R/(R+S)$. Note that the expected proportion of years with positive hybridization numbers is the same as in the crop rotation scenario considered in (16). We assume that the process is stationary. The hazard rate is then given by (see Appendix A.6)

$$H_n(q) = \frac{S}{R+S}\left(1 - e^{-m_0 p_0\left(1 - f_{I_1(0,n-1)}(q)\right)}\right). \tag{22}$$

Using the solution of $f_{I_1(0,n)}(q)$ from (6) and taking large $n$ leads to the asymptotic value:

$$H_\infty(q) = \frac{S}{R+S}\left[1 - e^{-m_0\, p_0 \beta_1(q)}\right]. \tag{23}$$

To examine the effects of autocorrelation, let $\kappa_{1,2} = \kappa_{2,1} = 1 - \kappa_{1,1} = 1 - \kappa_{2,2} = \alpha$. The environment is negatively autocorrelated if $\alpha > 0.5$, positively autocorrelated if $\alpha < 0.5$, and independent if $\alpha = 0.5$. The equations given in Appendix A.6 can be used to calculate the hazard rate for these models numerically . Figure 3a shows the resulting asymptotic hazard rate for different values of $\alpha$. As can be seen, there is not much difference between negatively autocorrelated or independent environments. The asymptotic hazard rate is much reduced, however, when there is a strong positive autocorrelation. With this choice of parameters, the probability of a year with hybridization is 1/2, and so the situation is comparable to a crop rotation scenario with $S = R$, as in Fig. 2(c). Note that the situation where $\alpha = 1$ corresponds to deterministic alternation between one-year periods with and without a positive hybridization probability. In this scenario, the hazard rate still approaches an asymptotic hazard rate because the process is initiated by the stationary-distribution of type-$(0,1)$ and type-$(0,2)$ individuals, as depicted in Fig. 3b. In a specific realisation, the hazard rate then oscillates as previously observed, which is also shown in Fig. 3b, where the process is initiated by a single type-$(0,1)$ individual.

## 6. Effects of bottlenecks

Until now we have considered the situation where all backcrossed generations are more fit than the wild type . However, often there is outbreeding depression, which implies that several backcrosses are needed before a fitness advantage is

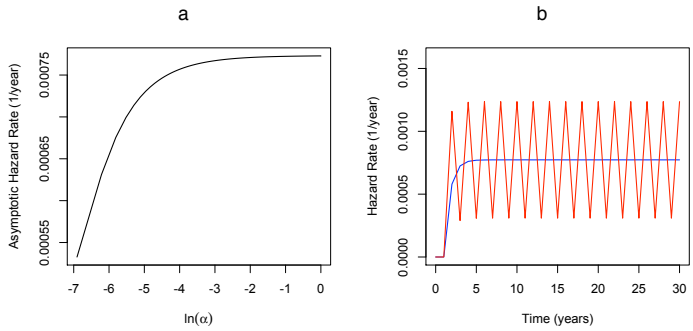


FIGURE 3. (a)The effect of autocorrelation on the asymptotic hazard rate when $k_{1,2} = k_{2,1} = \alpha = 1 - k_{2,2} = 1 - k_{1,1}$, $m_{0,1} = 50$, $m_{0,2} = 0$, and other parameter values as in Fig.2. The environment is positively autocorrelated when $\ln \alpha < \ln 2 (\approx -0.69)$ and negatively autocorrelated when $\ln \alpha > \ln 2$. Periods with and without positive hybridization probabilities alternate deterministically when $\ln \alpha = 0$. (b) Tthe hazard rate at $\alpha = 1$ when the process is started with a stationary distribution of type-$(0,1)$ and type-$(0,2)$ individuals (blue), and when the process is started with a single type-$(0,1)$ individual (red).

observed (e.g. Edmands, 2002). In this section we extend the model to account for such situations, and investigate effects of the length of the bottleneck on the hazard rate.

The generalized model involves $L+2$ ($L \in \mathbb{N}$) different types: types $0, 1, ..., L$, and type-$E$. Type-0 individuals are defined as before. The flowering probability of type-$i$ ($i \in \{1, 2, ..., L\}$) is denoted by $r_i$, the p.g.f. of their seed production by $G_i(s)$ and their seeds will produce type-$(i+1)$ adults. The survival probability of non-flowering type-$i$ individuals is $p_i$, and survivors remain of type $i$. The offspring of type-$L$ individuals will be of type-$E$. Type-$E$ individuals and $q$ are defined as in previous sections. The scheme is represented in Fig. 4.

The hazard rate in this scenario follows a similar method to the derivation in the previous case, but see Appendix A.7 for full details. Numerical solutions of the supremum of the hazard rate against $L$ are shown in Fig. 5a for the crop-rotation situation described in (16).

To further examine the effect of bottlenecks, we consider a Taylor approximation of the hazard rate around the point $q = 1$, for the case that plants are annual (i.e. $r_i = 1$ for $i = 1, 2, ..., L$). The resulting Taylor approximation is (see A.8 for details):

$$H_n(q) \approx \left( p_0 m_0 (n - L - 1) \prod_{i=1}^{L} p_0 m_i \right) (1 - q) \tag{24}$$

where $m_i$, $i = 1, 2, \ldots, L$, represents the average number of seeds produced by a type-$i$ individual.

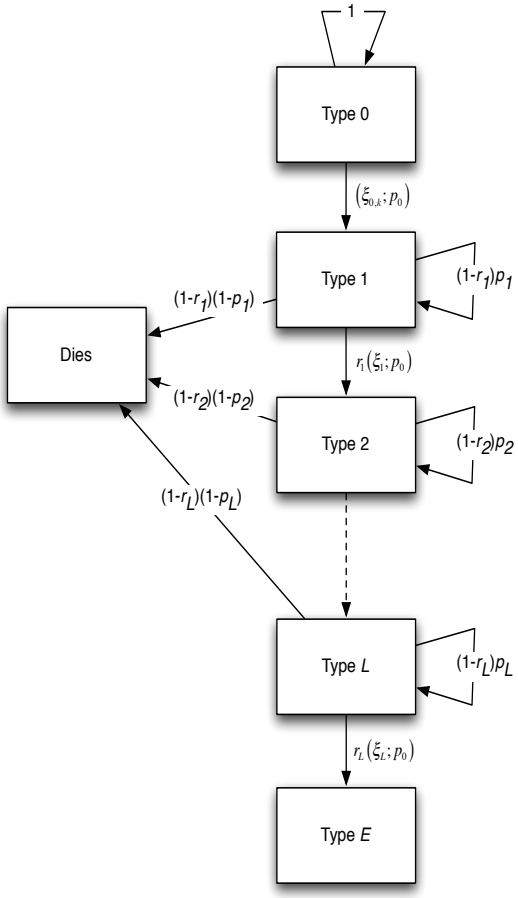


FIGURE 4. Schematic representation of the bottleneck model. $(\xi_i; p_0)$ represents the production of $\xi_i$ seeds $i \in (0, k) \cup \{1, 2, ..., L\}$, where each seed has a germination probability $p_0$. Each type-$E$ individual initiates a lineage which eventually becomes extinct with probability $q$.

When the values of $m_i$ are similar, this expression decreases geometrically with $L$, which corresponds to the shape observed in Fig. 5a.

Bottlenecks not only reduce the maximum hazard rate, but also induce a delay in the changes of the hazard rate in reaction to changes in crop cultivation. This is illustrated in Fig. 5b.

## 7. Discussion

In this paper we generalize our previous results on hazard rates of introgression (Ghosh and Haccou, 2010) to situations with time-varying hybridization. Whereas in our previous paper we considered a model with two age classes and a bottleneck of one generation, the present paper concerns situations without age dependence,

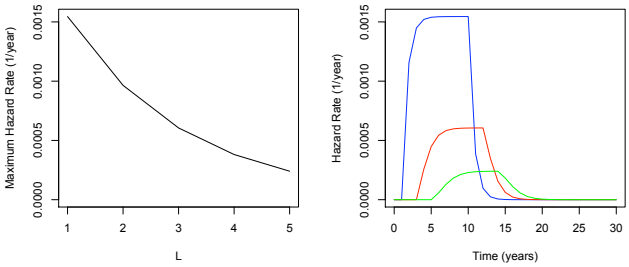


FIGURE 5. (a) The maximum hazard rate as a function of the length of the bottleneck $L$ for a crop rotation scenario with $R = S = 5$, $m_0 = 50$, $p_0 = 0.001$, $p_i = r_i = 0.5$, $m_i = 950$ for $i = 1, 2, ..., L$ and $q = 0.95$. (b) The hazard rate against time with hybridization as described in (12) with $S = 10$ and all other parameters as in (a). The behaviour for $L = 1$ (blue) $L = 3$ (red) and $L = 5$ (green) is shown.

and effects of extended bottlenecks. The general methodology that we present can be extended straightforwardly to other types of life histories. Furthermore, there are several general conclusions that are valid for a wide range of situations.

First of all, the results shed light on the meaning of the hazard rate as a measure of stochastic introgression rate, and its practical implications. As illustrated in this paper, hazard rates may increase and decrease in time, in relation to changes in the magnitude of hybridization rates. When the hybridization rate is high, the instantaneous risk of introgression events is also high. During such periods, increased vigilance is advisable, to prevent the successful establishment of crop genes in wild populations. When hybridization frequency drops, the hazard rate decreases, and accordingly, vigilance might be decreased. Our results show, however, that managers must take care not to let their guards down too soon, since increased fitness bottlenecks delay the changes in the hazard rate. This implies, for instance, that even after crop cultivation has been terminated for a considerable time, the risk of introgression events may still be quite high (see Fig. 5b), reaffirming a conclusion from Haygood et al. (2003).

The risk that introgression occurs is determined by the interaction between life history and fitness characteristics of hybrids, and crop management. As we illustrated, changes in gene flow induce changes in the level of the hazard rate. The speed at which such changes take place, as well as the magnitude of the hazard rate depends on life-history characteristics. For instance, increases in fitness bottlenecks not only cause a delay in adjustment of the hazard rate, but also decelerate the adjustments, and lower the maximum level. Furthermore, in all scenarios, the maximum level of the hazard rate is affected by the factor $\beta_1(q)$, which is determined by the fitness of the backcrosses (see (7)).

We examined the effect of several possible scenarios. With temporary crops, there is a positive probability that introgression does not occur, that depends on the duration of the crop cultivation. Furthermore, in this situation, the hazard

rate at a given time $x$ is nearly equal to the probability of an introgression event at that time, and thus provides a good approximation for the probability distribution (see e.g. Fig. 2b). This is a general result, that can be derived from the relation between the hazard rate and the time-distribution.

With crop rotation, the hazard rate becomes periodic, and fluctuations also occur in the time-distribution of introgression events (Fig. 2c and d). In such situations, a simpler measure of risk might sometimes be needed. One option is to use the hazard rate that in the long run would lead to the same introgression risk over a given period as the crop rotation scheme. This value is given in (21), and indicated in Fig. 2c. We refer to this value as the long-run average hazard rate. However, please note that it is not the same as the arithmetic time-average of the asymptotic hazard rate. From (21) it can be seen that the average risk level is determined by the proportion of years that crop cultivation occurs. Thus, the average hazard rate remains the same when $S$ and $R$ are multiplied by the same factor. For instance, alternating between one year 'on' and 'off' would in the long run give the same average hazard rate as alternating between, say, ten years 'on' and 'off'. Larger values of $S$ and $R$ would, however, lead to a larger amplitude of the fluctuations in the hazard rate. The magnitude of this effect can be calculated by means of (19) and (20). In situations with large fluctuations the use of the average hazard rate as a risk indicator might be misleading, since the maximum hazard rate is much higher than the average. This is illustrated in Fig. 2c. In such a situation, the time-distribution of introgression events corresponding to the average hazard rate is also radically different from the real one (see Fig. 2d).

Another possible way to quantify the risk is to use the long-run maximum hazard rate, which provides a conservative measure of risk. Figure 2d also shows the time-distribution of introgression events corresponding to the maximum hazard rate, illustrating that in an example with large amplitude of the hazard rate this might be a better risk measure.

We also derived methods to calculate the hazard rate in situations with randomly varying hybridization rates. As a specific example, we considered a situation where the environment alternates between two states, one with and one without hybridization, according to a Markov chain. In the absence of environmental autocorrelation, the hazard rate becomes constant in the long run, and an explicit expression can be derived. This value is given in (23), and corresponds to the arithmetic time-average of the asymptotic hazard rate in a deterministic crop rotation scheme with the same proportion of years of hybridization as the random environment. It can be shown that this value is lower than the long-run average hazard rate given in (21). Therefore, random variation in gene flow appears to reduce the probability that introgression occurs. This also appears to be true in autocorrelated environments, as illustrated in Fig.3. Positive autocorrelation reduces the hazard rate, whereas negative autocorrelation does not seem to have much effect. In any case, the long-run hazard rate is smaller than the long-run average for the deterministically alternating environment. Thus, we expect that hazard rates for deterministic scenarios provide conservative measures for introgression risk. This is a fortunate result, since in many situations there is likely to be random variation in gene flow, which is beyond control of management measures.

We examined several specific gene flow scenarios, to illustrate the methodology and its possibilities. For mathematical tractability, we used a relatively simple life-history and Poisson distributions for the numbers of hybrids. Our methods can readily be adjusted to examine other types of gene flow variation, more complicated life histories, and hybrid number distributions. In such cases, however, no explicit expressions for (asymptotic) hazard rates will be possible. Instead, numerical methods will have to be used, based on the adjusted equations. Such calculations generally do not take much time on a standard computer.

Other generalizations, which are the subject of ongoing research, include the introduction of time-inhomogeneity in backcross fitness, multi-locus genetics, and meta-population dynamics. Another type of generalization concerns small populations. As long as wild receptor populations are assumed to be large enough to exclude direct interactions between initial invaders, the approach that we used up to now, based on branching processes, can be applied. For small populations however, different methods need to be developed, based on density-dependent models (see (e.g Jagers and Klebaner, 2000)). This is another line of ongoing research.

The use of stochastic models in introgression studies is quite rare, although not completely absent (e.g. Haygood et al., 2004; Thompson et al., 2003). The general methodology for handling such models, and quantifying introgression timing events is, however, still in its infancy. The use of hazard rates is, in our opinion, an important step forward. Serra and Haccou (2007) introduced the concept of the hazard rate for studying branching processes with mutation, and Ghosh and Haccou (2010) were the first to use it in the context of introgression. The work presented here represents the next step of a research program that is aimed at developing a full-fledged toolbox for studying stochastic introgression processes. Such tools are indispensable in introgression risk management, since stochastic elements are inevitably present, and, furthermore, adding stochasticity changes the features of introgression processes considerably.

## Appendix A. Appendix

A.1. **Derivation of** (2)**.** Using (1) and the definition of p.g.f.'s we find:

$$\begin{aligned} f_{I_i(k,n)}(s) &= E\left[E\left[s^{I_i(k,n)}|Z_0^{(i)}(k+1), Z_1^{(i)}(k+1), Z_E^{(i)}(k+1)\right]\right] \\ &= E\left[E[s^{I_0(k+1,n)}]^{Z_0^{(i)}(k+1)}\ E[s^{I_1(k+1,n)}]^{Z_1^{(i)}(k+1)} E[s]^{Z_E^{(i)}(k+1)}\right] \\ &= E\left[f_{I_0(k+1,n)}(s)^{Z_0^{(i)}(k+1)} f_{I_1(k+1,n)}(s)^{Z_1^{(i)}(k+1)} s^{Z_E^{(i)}(k+1)}\right] \end{aligned} \tag{A.1}$$

We can manipulate (A.1) as above because the individual lineages are independent of each other, and individuals of the same type have identical offspring distributions.

Now we introduce the joint p.g.f of the reproduction distribution of a type-$i$ individual belonging to a year $k$ which, for $i \in \{0,1\}$ and $k \geq 0$, is defined as

$$F_i\left(k;(s_0,s_1,s_E)\right) = E\left[s_0^{Z_0^{(i)}(k+1)} s_1^{Z_1^{(i)}(k+1)} s_E^{Z_E^{(i)}(k+1)}\right] \tag{A.2}$$

for $(s_0, s_1, s_E) \in [0,1]^3$.

Putting (A.1) and (A.2) together, we find that

$$f_{I_i(k,n)}(s) = F_i(k;(f_{I_0(k+1,n)}(s), f_{I_1(k+1,n)}(s), s)) \tag{A.3}$$

In our specific model, we have the following reproduction laws:

$$F_0(k;(s_0,s_1,s_E)) = s_0 G_0(k; p_0 s_1 + (1-p_0)) \tag{A.4}$$

$$F_1(k;(s_0,s_1,s_E)) = (1-r_1)(1-p_1) + (1-r_1)p_1 s_1 + r_1 G_1(p_0 s_E + 1 - p_0). \tag{A.5}$$

Substituting (A.5) and (A.4) into (A.3) gives (2).

A.2. **Derivation of** (6)**.** Since the population initiated by a type-1 individual is time-homogeneous, $f_{I_1(k,n)}(s) = f_{I_1(0,n-k)}(s)$. Using this in the second equation of (2) results in:

$$f_{I_1(0,n-k)}(s) = (1-r_1)(1-p_1) + (1-r_1)p_1 f_{I_1(0,n-k-1)}(s) + r_1 G_1(p_0 s + 1 - p_0) \tag{A.6}$$

Introducing $b_1 = (1-r_1)p_1$ and $a_1(s) = (1-r_1)(1-p_1) + r_1 G_1(p_0 s + 1 - p_0)$, allowing $k = 0$, this can be rewritten as follows:

$$\begin{aligned} f_{I_1(0,n)}(s) &= a_1(s) + b_1 f_{I_1(0,n-1)}(s) \\ &= a_1(s) + b_1\left(a_1(s) + b_1 f_{I_1(0,n-2)}(s)\right) \\ &= ... \\ &= b_1^n + a_1(s)\sum_{i=0}^{n-1} b_1^i. \end{aligned} \tag{A.7}$$

Computing the geometric sum above, and taking the quantities defined in (6) gives the required result.

A.3. **Derivation of** (8)**.** Deriving (8) follows from repeating equation (2) in the following way:

$$\begin{aligned} f_{I_0(0,n)}(s) &= f_{I_0(1,n)}(s) G_0(0; p_0 f_{I_1(1,n)}(s) + 1 - p_0) \\ &= f_{I_0(2,n)}(s) G_0(1; p_0 f_{I_1(2,n)} + 1 - p_0) G_0(0; p_0 f_{I_1(1,n)}(s) + 1 - p_0) \\ &\vdots \\ &= \prod_{j=1}^{n-1} G_0(j-1; p_0 f_{I_1(0,n-j)}(s) + 1 - p_0) \end{aligned} \tag{A.8}$$

The expression in (8) follows from substituting (A.8) into (5).

A.4. **Derivation of** (18)**.** Substituting (16) into (10) gives the hazard rate. During the $(v+1)^{th}$ period that hybridization is introduced, i.e. if $v(R+S)+2 \leq n < v(R+S)+S+2$, the following holds:

$$H_n(q) = 1 - e^{-m_0 p_0 \beta_1(q) \left( 1 - b_1^{n-(1+v(S+R))} + b_1^{n-(S+1)} (1-b_1^S) \left( \frac{1-b_1^{v(R+S)}}{b_1^{(v-1)(R+S)} \left(1-b_1^{R+S}\right)} \right) \right)} \tag{A.9}$$

and for the $(v+1)^{th}$ period that hybridization is stopped, i.e. if $v(R+S)+S+2 \leq n < (v+1)(R+S)+2$,

$$H_n(q) = 1 - e^{-m_0 p_0 \beta_1(q) b_1^{n-(S+1)} \left(1-b_1^S\right) \left( \frac{1-b_1^{(v+1)(R+S)}}{b_1^{v(R+S)} \left(1-b_1^{R+S}\right)} \right)} \tag{A.10}$$

and, as in (10), the hazard rate equals zero for $n \in \{0, 1\}$. Substituting (17) into (A.9) leads to the following for $0 \leq k < S$ :

$$H_{v(R+S)+2+k}(q) = 1 - e^{-m_0 p_0 \beta_1(q) \left( 1 - b_1^{k+1} + b_1^{v(R+S)+k+1-S} \left(1-b_1^S\right) \left( \frac{1-b_1^{v(R+S)}}{b_1^{(v-1)(R+S)} \left(1-b_1^{(R+S)}\right)} \right) \right)} \tag{A.11}$$

and substituting (17) into (A.10) leads to, for $S \leq k < S+R$:

$$H_{v(R+S)+2+k}(q) = 1 - e^{-m_0 p_0 \beta_1(q) b_1^{v(R+S)+k+1-S} \left(1-b_1^S\right) \left( \frac{1-b_1^{(v+1)(R+S)}}{b_1^{v(R+S)} \left(1-b_1^{(R+S)}\right)} \right)}. \tag{A.12}$$

To reach the asymptotic behaviour described in (18), take $v \to \infty$ in both (A.11) and (A.12).

A.5. **Derivation of** (21)**.** First, note that the survival function of $T$ and the hazard rate are related as follows. For any $t \in [0, +\infty)$:

$$P[T > t] = \prod_{j \in \mathbb{N}_0 \,:\, j \leq t} \left(1 - H_j\left(q\right)\right). \tag{A.13}$$

Define the sequence $\{c_n, n \in \mathbb{N}_0\}$:

$$c_n = \frac{P[T > n+R+S]}{P[T > n]}. \tag{A.14}$$

The use of (A.8) with (4) and (A.14), gives:

$$\begin{aligned} c_n &= \frac{f_{I_0(0,n+R+S)}(q)}{f_{I_0(0,n)}(q)} \\ &= \frac{\prod_{i=1}^{n+R+S-1} G_0(i-1; p_0\, f_{I_1(0,n+R+S-i)}(q) + 1 - p_0)}{\prod_{i=1}^{n-1} G_0(i-1; p_0\, f_{I_1(0,n-i)}(q) + 1 - p_0)} \\ &= \frac{\prod_{i=1}^{R+S} e^{-p_0 m_0 (i-1)(1-f_{I_1(0,n+R+S-i)}(q))} \prod_{i=R+S+1}^{n+R+S-1} e^{-p_0 m_0 (i-1)(1-f_{I_1(0,n+R+S-i)}(q))}}{\prod_{i=1}^{n-1} e^{-p_0 m_0 (i-1)(1-f_{I_1(0,n-i)}(q))}} \\ &= e^{-p_0\, m_0 \sum_{i=1}^{S}(1 - f_{I_1(0,n+R+S-i)}(q))} \end{aligned} \tag{A.15}$$

Note how the second product in the numerator is identical to the denominator. This is a result of the periodicity of the hybridization rate in (16). Also, note that for $S+1 \leq i \leq R+S$, $m(i) = 0$, which is used to reduce the number of terms in the sum.

When $n \to \infty$, $c_n$ converges to

$$C = e^{-p_0 m_0 S \beta_1(q)}. \tag{A.16}$$

Thus, in the long run, a process with a constant hazard rate, $\lambda$, and such that $\lim_{n\to\infty} \frac{P[T>n+R+S]}{P[T>n]} = C$, would have the same probability of an introgression event occurring within a period from $n$ to $n+R+S$, with sufficiently large $n$. Using (A.13) and (A.16) we find that $\lambda$ must satisfy

$$\lim_{n\to\infty} \prod_{i=n+1}^{n+R+S} (1-\lambda) = C, \tag{A.17}$$

and the required result follows by combining (A.16) and (A.17) and solving for $\lambda$.

A.6. **Derivation of** (22)**.** Take the definitions of $f_{I_i(k,n)}(s)$, $I_i(k,n)$ and $Z_m^{(i)}$ as before, but extend it to include $i = (0,1)$ and $(0,2)$. As before, a joint p.g.f. of the offspring distribution of a single type-$i$ ($i = (0,1), (0,2), 1, E$) is defined:

$$F_i\left(k; (s_{0,1}, s_{0,2}, s_1, s_E)\right) = E\left[ s_{0,1}^{Z_{0,1}^{(i)}(k+1)} s_{0,2}^{Z_{0,2}^{(i)}(k+1)} s_1^{Z_1^{(i)}(k+1)} s_E^{Z_E^{(i)}(k+1)} \right] \tag{A.18}$$

Then, following the same methodology established in A.1, we get:

$$f_{I_i(k,n)}(s) = F_i(k; (f_{I_{0,1}(k+1,n)}(s), f_{I_{0,2}(k+1,n)}(s), f_{I_1(k+1,n)}(s), s)) \tag{A.19}$$

Following further the methodology in A.1, the following recursive relationships hold:

$$\begin{aligned} f_{I_{0,1}(0,n-k)}(s) =& G_{0,1}\left(p_0 f_{I_1(0,n-k-1)}(s) + 1 - p_0\right) \times \\ & \left(\kappa_{1,1} f_{I_{0,1}(0,n-k-1)}(s) + \kappa_{1,2} f_{I_{0,2}(0,n-k-1)}(s)\right) \qquad \text{(A.20)} \\ f_{I_{0,2}(0,n-k)}(s) =& G_{0,2}\left(p_0 f_{I_1(0,n-k-1)}(s) + 1 - p_0\right) \times \\ & \left(\kappa_{2,1} f_{I_{0,1}(0,n-k-1)}(s) + \kappa_{2,2} f_{I_{0,2}(0,n-k-1)}(s)\right) \qquad \text{(A.21)} \end{aligned}$$

where the simplifying expression $f_{I_i(k,n)}(s) = f_{I_i(0,n-k)}(s)$ has been applied. Using the forms of $G_{0,1}(s)$ and $G_{0,2}(s)$ as specified in section 5, and setting $k = 0$, gives:

$$\begin{aligned} f_{I_{0,1}(0,n)}(s) &= e^{-m_0 p_0\left(1 - f_{I_1(0,n-1)}(s)\right)} \times \\ &\quad \left(\kappa_{1,1} f_{I_{0,1}(0,n-1)}(s) + \kappa_{1,2} f_{I_{0,2}(0,n-1)}(s)\right) \\ f_{I_{0,2}(0,n)}(s) &= \kappa_{2,1} f_{I_{0,1}(0,n-1)}(s) + \kappa_{2,2} f_{I_{0,2}(0,n-1)}(s) \qquad \text{(A.22)} \end{aligned}$$

Since the environmental process is stationary:

$$P\left(T > n\right) = \frac{\kappa_{2,1}}{\kappa_{1,2} + \kappa_{2,1}} f_{I_{0,1}(0,n)}(q) + \frac{\kappa_{1,2}}{\kappa_{1,2} + \kappa_{2,1}} f_{I_{0,2}(0,n)}(q), \qquad \text{(A.23)}$$

and the hazard rate can be calculated from this.

For the considered analog of the deterministic process without autocorrelation, $f_{I_{0,1}(0,n)}(s) = e^{-m_0 p_0\left(1 - f_{I_1(0,n-1)}(s)\right)} f_{I_{0,2}(0,n)}(s)$ . Using (5) (A.23) and (A.22) then gives the required result.

A.7. **Derivation of the hazard rate in the bottleneck scenario.** We start by defining the random variable $I_i(k,n)$ as before, except with $i \in \{0,1,\ldots,L\}$. Also, we define p.g.f.'s, $f_{I_i(k,n,)}(s)$, of these random variables in the same way as previously done.

Since an individual belonging to a generation greater than $n$ can produce no type-$E$ individuals before $n$, write the following for any $i \in \{0,1,\ldots,L\}$,

$$I_i(k,n) = 0, \quad \text{if } k \geq n. \qquad \text{(A.24)}$$

Let us now turn to the case $k < n$. For a fixed $i \in \{0,\ldots,L\}$, and a general scenario, where individuals can have offspring of any type, the following decomposition holds

$$I_i(k,n) = Z_E^{(i)}(k+1) + \sum_{m=0}^{L} \sum_{j=1}^{Z_m^{(i)}(k+1)} I_m^{(j)}(k+1,n)\,, \qquad \text{(A.25)}$$

where the random variables

$$Z_0^{(i)}(k+1),\ \ Z_1^{(i)}(k+1),\ \ldots,\ Z_L^{(i)}(k+1),\ \ Z_E^{(i)}(k+1)$$

represent the number of offspring of types $0,1,\ldots,L,E$, respectively, that the initial type $i$ produced. Also, as the notation suggests, the random variables

$$I_0^{(j)}(k+1,n),\ \ j = 1,\ldots,Z_0^{(i)}(k+1),$$

represent the number of type-$E$ individuals with non-type-$E$ parents, appearing up to and including year $n$, in the line of descent of the $j^{th}$ type-0 offspring of the

initial type-$i$ individual. Notice that, since the initial type-$i$ individual belongs to year $k$, its offspring belongs to year $k+1$. The random variables

$$\begin{array}{ll} I_1^{(j)}(k+1,n), & j=1,\ldots,Z_1^{(i)}(k+1), \\ I_2^{(j)}(k+1,n), & j=1,\ldots,Z_2^{(i)}(k+1), \\ & \vdots \\ I_L^{(j)}(k+1,n) & j=1,\ldots,Z_L^{(i)}(k+1), \end{array}$$

are defined in a analogous way, but now for the type-1, type-2, ..., type-$L$, respectively, offspring of the initial type-$i$ individual.

First manipulate the generating functions of (A.25) as follows:

$$\begin{aligned} f_{I_i(k,n)}(s) &= E\left[E\left[s^{I_i(k,n)}|Z_0^{(i)}(k+1), Z_1^{(i)}(k+1),\ \ldots, Z_L^{(i)}(k+1), Z_E^{(i)}(k+1)\right]\right] \\ &= E\left[f_{I_0(k+1,n)}(s)^{Z_0^{(i)}(k+1)} f_{I_1(k+1,n)}(s)^{Z_1^{(i)}(k+1)}\ \ldots\ f_{I_L(k+1,n)}(s)^{Z_L^{(i)}(k+1)} s^{Z_E^{(i)}(k+1)}\right] \end{aligned} \tag{A.26}$$

We can manipulate (A.26) as above because the individual lineages are independent of each other, and individuals of the same type have identical offspring distributions.

Introduce the joint p.g.f of the reproduction distribution of a type-$i$ individual belonging to a year $k$ which, for $i \in \{0,1,\ldots,L\}$ and $k \geq 0$, is defined as

$$F_i\left(k;(s_0,s_1,\ldots,s_L,s_E)\right) = E[s_0^{Z_0^{(i)}(k+1)} s_1^{Z_1^{(i)}(k+1)} \ldots s_L^{Z_L^{(i)}(k+1)} s_E^{Z_E^{(i)}(k+1)}] \tag{A.27}$$

for $(s_0,s_1,\ldots,s_L,s_E) \in [0,1]^{L+2}$.

Putting (A.26) and (A.27) together, we find that

$$f_{I_i(k,n)}(s) = F_i(k;(f_{I_0(k+1,n)}(s), f_{I_1(k+1,n)}(s),\ldots,f_{I_L(k+1,n)}(s),s)) \tag{A.28}$$

In our specific model, we have the following assumptions regarding the reproduction:

- the reproduction law of a type 0 individual depends on the year number and the corresponding p.g.f. is given by

$$F_0(k;(s_0,s_1,\ldots,s_L,s_E)) = s_0 G_0(k;p_0 s_1 + (1-p_0)) \tag{A.29}$$

- for a type $i$ individual, with $i \in \{1,\ldots,L\}$, the reproduction law does not depend on the year number and the corresponding p.g.f. is given by

$$\begin{aligned} F_i(k;(s_0,s_1,\ldots,s_i,s_{i+1},\ldots,s_L,s_E)) &= (1-r_i)(1-p_i) + (1-r_i)p_i s_i \\ &+ r_i G_i(p_0 s_{i+1} + 1 - p_0) \end{aligned} \tag{A.30}$$

with $s_{L+1} \equiv s_E$. The fact that the reproduction law of these individuals is independent of time implies that

$$f_{I_i(k,n)}(s) = f_{I_i(0,n-k)}(s).$$

This relation will be used more or less explicitly in the following calculations.

The use of (A.30) and (A.28) with $i = L$, gives

$$f_{I_L(0,n)}(s) = (1-r_L)(1-p_L) + (1-r_L)\,p_L\,f_{I_L(0,n-1)}(s) + r_L\,G_L(p_0 s + 1 - p_0).$$

The use of initial condition $f_{I_L(0,0)}(s) = 1$ results in the following for any $n \geq 0$, which is :

$$f_{I_L(0,n)}(s) = 1 - \beta_L(s) + \beta_L(s)\, b_L^n, \tag{A.31}$$

with

$$b_L = (1 - r_L)\, p_L \quad \text{and} \quad \beta_L(s) = \frac{r_L\,(1 - G_L(p_0 s + 1 - p_0))}{1 - b_L}. \tag{A.32}$$

The calculation of (A.31) above follows the same reasoning shown in Appendix A.2.

Now that we can calculate the p.g.f.'s of $I_L(0, n)$, we proceed by finding expressions for the p.g.f.'s of $I_i(0, n)$ for $i = 0, 1, \ldots L - 1$.

Note that, in the line of descent of a single type-$i$ individual belonging to year 0, new type-$E$ individuals can only appear after $L - i + 1$ years (this is intuitively clear from Fig. 4). Hence, for $i \in \{1, \ldots, L - 1\}$,

$$f_{I_i(0,1)}(s) = f_{I_i(0,2)}(s) = \ldots = f_{I_i(0,L-i)}(s) = 1.$$

Now, for $n > L - i$, the use of (A.30) and (A.28), gives

$$f_{I_i(0,n)}(s) = (1-r_i)(1-p_i)+r_i G_i(p_0 f_{I_{i+1}(0,n-1)}(s)+1-p_0)+(1-r_i)p_i f_{I_i(0,n-1)}(s).$$

Repeating the procedure gives

$$\begin{aligned} f_{I_i(0,n)}(s) &= [(1 - r_i)p_i]^{n-(L-i)} + (1 - p_i) \sum_{j=1}^{n-(L-i)} (1 - r_i)^j\, p_i^{j-1} \\ &+ \sum_{j=1}^{n-(L-i)} r_i[(1 - r_i)p_i]^{j-1} G_i(p_0\, f_{I_{i+1}(0,n-j)}(s) + 1 - p_0). \end{aligned}$$

Computing the sums above gives us the following p.g.f.'s:

$$f_{I_i(0,n)}(s) = 1 - \alpha_i + \alpha_i\, b_i^{n-(L-i)} + r_i \sum_{k=L-i}^{n-1} b_i^{n-k-1} G_i(p_0\, f_{I_{i+1}(0,k)}(s) + 1 - p_0), \tag{A.33}$$

where

$$b_i = (1 - r_i)p_i \quad \text{and} \quad \alpha_i = \frac{r_i}{1 - b_i}. \tag{A.34}$$

We have $f_{I_0(0,n)}(s) = 1$ for $n \leq L$, since a type-0 individual requires at least $L$ generations to produce a type-E individual. For $n > L$ we combine (A.29) and (A.28) to give:

$$f_{I_0(0,n)}(s) = \prod_{j=1}^{n-L} G_0(j - 1; p_0 f_{I_1(0,n-j)}(s) + 1 - p_0) \tag{A.35}$$

which can be calculated using (A.33) and (A.31).

The use of (A.35) and noting that, as before, $P(T > n) = f_{I_0(0,n)}(q)$ yields the hazard rate:

$$H_n(q) = \begin{cases} 0 & \text{if} \quad 0 \leq n \leq L \\ 1 - \dfrac{\prod_{j=1}^{n-L} G_0\left(j-1; p_0 f_{I_1(0,n-j)}(q) + 1 - p_0\right)}{\prod_{j=1}^{n-1-L} G_0\left(j-1; p_0 f_{I_1(0,n-1-j)}(q) + 1 - p_0\right)} & \text{if} \quad n \geq L+1. \end{cases} \tag{A.36}$$

A.8. **Derivation of** (24)**.** Taking $r_1 = 1$ in (A.31) to (A.34) gives:

$$f_{I_L(0,n)}(s) = 1 - \beta_L(s) \tag{A.37}$$

$$f_{I_i(0,n)}(s) = G_i\left(p_0 f_{I_{i+1}(0,n-1)}(s) + 1 - p_0\right) \tag{A.38}$$

where $i = 1, 2, ..., L-1$. Differentiating these expressions with respect to $s$ and evaluating the results at the point $s = 1$ gives:

$$\begin{aligned} f'_{I_L(0,n)}(1) &= p_0 m_L \\ f'_{I_i(0,n)}(1) &= p_0 m_i f'_{I_{i+1}(0,n-1)}(1) \end{aligned} \tag{A.39}$$

where we have used the fact that the derivative of a p.g.f. evaluated at one is the mean of the random variable.

Taking logarithms in (A.35) and differentiating at $s = 1$ yields the following expression:

$$\begin{aligned} f'_{I_0(0,n)}(1) &= \sum_{j=1}^{n-L} p_0 m_0 (j-1) f'_{I_1(0,n-j)}(1) \\ &= \sum_{j=1}^{n-L} p_0 m_0 (j-1) p_0^L \prod_{i=1}^{L} m_i \end{aligned} \tag{A.40}$$

where the last equality uses the expressions in (A.39).

Consider the representation of the hazard rate in (5). It is apparent that the constant-term in the Taylor approximation will be zero, due to the fact that p.g.f.'s evaluated at one are one. Taking the derivative of (5) around one yields:

$$H'_n(1) = f'_{I_0(0,n-1)}(1) - f'_{I_0(0,n)}(1). \tag{A.41}$$

Using the above with (A.40) gives the required result.

## References

Bos, P.M.J., Jan, P-J, van Raaij, M.T.M., 2004. Risk assessment of peak exposure to genotoxic carcinogens: a pragmatic approach. Toxicol. Lett. 151, 43-50.

Demon, I., Haccou, P., van den Bosch, F., 2007. Introgression of resistance genes between population: A model study of insecticide resistance in Bemisia tabaci. Theor. Popul. Biol. 72, 292-304.

Edmands, S., 2002. Does parental divergence predict reproductive compatibility? Trends Ecol. Evol. 17, 520-527.

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.

Ghosh, A., Haccou, P., 2010. Quantifying stochastic introgression processes with hazard rates. Theor. Popul. Biol. 77, 171-180.

Haccou, P., Jagers, P., Vatutin, V.A., 2005. Branching Processes: Variation Growth and Extinction of Populations. Cambridge University Press, Cambridge.

Hails, R.S., Morley, K., 2005. Genes invading new populations: A risk assessment perspective. Trends Ecol. Evol. 20, 245-252.

Haygood, R., Ives, A.R., Andow, D.A., 2003. Consequences of recurrent gene flow from crops to wild relatives. Proc. Biol. Sci. 270, 1879-1896.

Haygood, R., Ives, A.R., Andow, D.A., 2004. Population genetics of transgene containment. Ecol. Lett. 7, 213-220.

Jagers, P., Klebaner, F.C., 2000. Population-size-dependent and age-dependent branching processes. Stoch. Proc. Appl. 87, 235-254.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure Time Data, 2nd ed., John Wiley & Sons, New York.

Kelly, C.K., Bowler, M.J., Breden, F., Fenner, M., Poppy, G.M., 2005. An analytical model assissing the potential threat to natural habitats from insect resistance transgenes. Proc. Biol. Sci. 272, 1759-1767.

Maan, S.S, 1987. Interspecific and intergeneric hybridisation in wheat. In: Heyne, E.G. (Ed.), Wheat and Wheat Improvement. ASA, CSSA and SSSA, Madison, pp. 453-461.

Michor, F., Nowak, M.A., Iwasa, Y., 2006. Stochastic dynamics of metastatis formation. J. Theoret. Biol. 240, 521-530.

Reluga, T., Meza, R., Walton, D.B., Galvani, A.P., 2007. Reservoir interactions and disease emergence. Theor. Popul. Biol. 72, 400-408.

Riesberg, L.H., Wendel, J.F., 1993. Introgression and its consequences in plants. In: Harrison, R.G. (Ed.), Hybrid Zones and the Evolutionary Process. Oxford University Press, Oxford, pp. 70-109.

Serra, M.C., Haccou, P., 2007. Dynamics of escape mutants. Theor. Popul. Biol. 72, 167-178.

Snow, A.A., Andersen, B., Jorgensen, R.B., 1999. Costs of transgenic herbicide resistance introgressed from Brassica napus into weed B. Rapa. Mol. Ecol. 8, 605-615.

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.

Welliver, R., 2009. The relationship of meteorological conditions to the epidemic activity of respiratory syncytial virus. Paediatr. Respir. Rev., 10 (2009) 6-8.

# 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

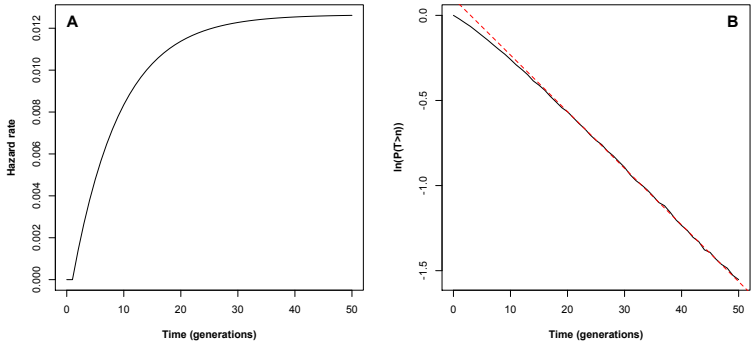


FIGURE 1. Calculation of the hazard rate for both models. A) The hazard rate plotted against time, as calculated by the branching process method. The asymptote reached corresponds to the value given in (3). B) A sample output from the simulation based method. $\ln P(T > n)$ plotted against time (black 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.

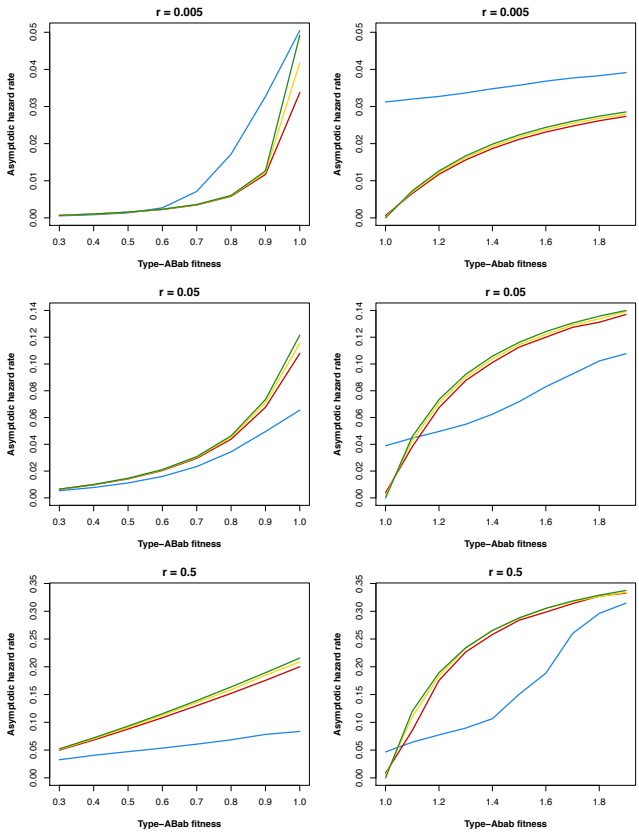


FIGURE 2. The asymptotic hazard rate as a function of the fitness of type-*ABab* individuals (left), and against the fitness of type-*Abab* individuals (right hand), for three different recombination rates. For the left column, $m_{ABab}$ is varied from 0.3 to 1.0, with $w_{Abab}$ set to 1.2; for the 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.

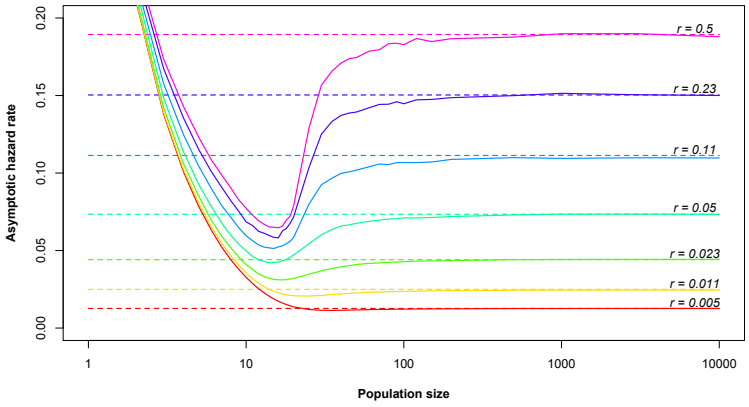


FIGURE 3. The asymptotic hazard rate plotted against population size (on a logarithmic scale) for $w_{ABab} = 0.9$, $w_{Abab} = 1.2$, $m = 1$, and various recombination rates in different colours as indicated in the plot. The branching process results, which are independent of the population size, are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

# CHAPTER 5: QUANTIFYING STOCHASTIC INTROGRESSION PROCESSES IN RANDOM ENVIRONMENTS WITH HAZARD RATES

To be submitted

ABSTRACT

Introgression is the permanent incorporation of genes from the genome of one population into another. Fears that genetically modified genes might introgress from crop populations into their wild relatives has prompted many theoretical attempts to quantify the risk of introgression. Previous studies have found that stochasticity in number of offspring, hybridization, and environment are important aspects of introgression risk, but so far studies have considered these factors separately, and they have not yet been combined into one framework. In this paper we develop such a framework. In previous papers we introduced a a measure of risk known as the hazard rate of introgression, that accurately takes demographic stochasticity into account. Here, we extend the methodology to incorporate random temporal environmental variation. We find that introgression risk varies much in time, and in some periods it can be much enhanced in such environments. Furthermore, effects of plant life history parameters, such as flowering and survival probabilities, depend on environmental variation.

## 1. Introduction

The permanent incorporation of genes from the genome of one population into another, a process known as introgression, is a topical area of research which has garnered much attention due to fears that transgenes might enter wild populations from crop populations, e.g. Kwit et al. (2011), Ellstrand et al. (1999) and Hails and Morley (2005). Potential consequences of introgression are, for example, the displacement of local species (as described in Huxel (1999)) or the creation of so-called super weeds via the transfer of herbicide resistance to wild individuals (e.g. Reichmann et al. (2006)).

A key factor in modeling introgression risks is the randomness of the environment. Davis et al. (1999) and Thompson et al. (2003) included environmental stochasticity in their models, but they did not consider demographic stochasticity. Ghosh and Haccou (2010) showed that demographic stochasticity is an important factor that should not be disregarded, especially with repeated outcrossing. The combined effects of environmental and demographic stochasticity have, until now, hardly been examined at all. In a previous paper (Ghosh et al. in press) we initiated such a study, by looking at stochastic changes in outcrossing rates. This can be caused, e.g., by variation in weather conditions. In the present paper we further generalize the methods to include randomness in other environmental conditions, which may affect the survival and reproduction of hybrids and further backcrosses. This type of environmental randomness is technically more difficult

to include, since every environmental change influences the complete future of an introgression process.

Ghosh and Haccou (2010), were the first to propose the hazard rate as a measure of introgression risk when there are repeated invasions. This measure is defined as the probability per time unit that the first introgressed lineage is initiated. Hazard rates are commonplace in medical statistics and behavior analysis (e.g. Kalbfleisch and Prentice (2002) and Haccou and Meelis (1994)), but they provide an intuitive measure to quantifying invasion risks too. Ghosh and Haccou (2010) calculated hazard rates of introgression by considering the repeated invasion of a gene conferring some fitness advantage into a large wild population in a temporally homogeneous environment. We showed that such environments lead to a monotonically increasing hazard rate that converges to some asymptote. Ghosh et al. (In press) demonstrated that deterministic temporal inhomogeneities can lead to a non-monotonic hazard rate, and thus that introgression risks can be higher at some times than at others. Generalizing the approaches first presented in Ghosh and Haccou (2010) to random environments involves incorporating theory on branching processes in random environments. Specifically, we will make use of the numerical methods for determining extinction probabilities in random environments. which were developed by Haccou and Iwasa (1996) and Haccou and Vatutin (2003). The methods presented in this paper may also be applied in contexts other than plant gene introgression that concern invasion with repeated immigration, for example in the study of invasive species, or epidemiological problems.

## 2. The model

We consider the model used in Ghosh et al. (In press) as an example. The methodology can straightforwardly be generalized to more complex ecological and life history settings. The model is in discrete time, with one time unit corresponding to one year. Plants are assumed to be monocarpic (i.e. they flower once then die), and there is no age-dependence in the life-history parameters. We assume that the recipient wild population is large and stable. A random number of hybrid seeds is produced each year, due to pollen flow from a neighboring crop. Seeds might germinate with some probability at the beginning of a year, and can flower in the same year. Whereas, previously, these germination and flowering probabilities were assumed to be fixed, in the present paper they may vary randomly in time.

As before, we will incorporate hybridization into the model by means of an artificial type, called type-0. There is always one single type-0 individual that produces a stochastic number of hybrid seeds each year. We will refer to hybrids as type-1 individuals.

The hybrids can backcross with the wild population, and subsequent backcrossses can backcross again with wild plants. All backcrosses are assumed to be equivalent, i.e. we assume that there are no fitness effects of further backcrossing after BC1. Also, we assume that there are no relevant genetical differences between hybrid or backcrossed individuals of the same generation. In Ghosh et al. (Submitted) we present methods for incorporating more realistic genetical mechanisms and fitness effects. Backcrossed individuals are called type-$E$ (as in escape type) individuals.

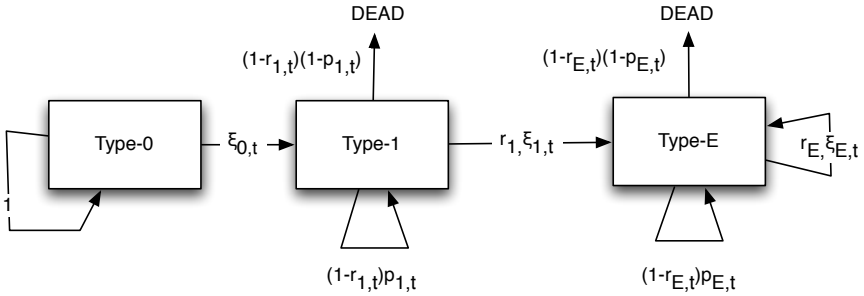


FIGURE 1. A schematic representation of the plant types used in the model.

We assume that the wild population is large relative to the initial numbers of hybrids and backcrossed individuals, so that the probability of interaction between individuals carrying crop genes is negligibly small. Consequently, individuals carrying crop genes interact solely with wild type individuals. This implies that a branching process model can be used to study the invasion dynamics.

In year $t$, a type-0 individual produces one individual of type-0 and a random number, $\xi_{0,t}$, of F1 hybrid seeds, each of which germinates with a probability of $p_{0,t}$ to become a type-1 plant of the next generation. Type-1 individuals at a time-$t$ flower with probability $r_{1,t}$ to produce a random number of backcrossed seeds, through both male and female functions. Each of these seeds germinates with probability $p_{0,t}$ as before, to make type-$E$ individuals belonging to the next generation. In the case that a type-1 individual does not flower, it may survive with a probability $p_{1,t}$. Similar dynamics hold for type-$E$ individuals (see Fig. 1). The flowering and survival probabilities, as well as the distributions of the offspring numbers, may depend on the environmental state.

## 3. Derivation of the hazard rate

To calculate the hazard rate, we first determine the sequence of extinction probabilities of a process initiated by a single individual at time $t$, by means of the methods that were developed by Haccou and Iwasa (1996) and Haccou and Vatutin (2003). Define $Q_{t,n}$ to be the probability that the lineage initiated by a single type-$E$ individual at a time $t$ becomes extinct at or before time $n$, conditioned on all environment states. This leads to the following expression for $t < n$:

$$Q_{t,n} = (1-r_{E,t})(1-p_{E,t}) + (1-r_{E,t})p_{E,t}Q_{t+1,n} + r_{E,t}G_E\left(t; p_{0,t}Q_{t+1,n} + 1 - p_{0,t}\right), \tag{1}$$

with $Q_{n,n} = 0$. Note that $G_i(t;s)$ $(i \in \{0,1,E\})$ represents the probability generating function (p.g.f.) of $\xi_{i,t}$. $Q_{t,n}$ is calculated for all $t$ and $n$ for the simulated sequence of environmental states. For large $n$ this gives the asymptotic extinction probability of the lineage initiated by a single type-$E$ individual belonging to generation $t$, which we will write as $Q_t$.

Now define $I_i(k,n)$ $(i \in \{0,1\})$ to be a random vector of length $n-k$, where the $j$th $(j \in \{1,2,...,n-k\})$ element represents the total number of type-$E$ individuals

belonging to generation $k+j$ produced in the lineage initiated by a single type-$i$ individual belonging to generation-$k$.

We now introduce the joint p.g.f. of $I_i(k,n)$, which is called $F_{I_i(k,n)}(s_{k+1}, s_{k+2}, s_{k+3}..., s_n)$ and is defined as:

$$F_{I_i(k,n)}(s_{k+1}, s_{k+2}, ..., s_n) = E\left[s_{k+1}^{Z_{k,k+1}^{(i)}} s_{k+2}^{Z_{k,k+2}^{(i)}} s_{k+3}^{Z_{k,k+3}^{(i)}} ... s_n^{Z_{k,n}^{(i)}} | \text{environmental sequence}\right], \tag{2}$$

where $Z_{k,j}^{(i)}$ represents the $j$th element of the vector $I_i(k,n)$.

We define the time of an introgression event, $T$, as the time where the first type-$E$ individual appears, whose lineage escapes extinction. Using the value of $Q_t$ already calculated, along with the definition in (2), we find:

$$P(T > n | \text{environmental sequence}) = F_{I_0(0,n)}(Q_1, Q_2, ..., Q_n), \tag{3}$$

since introgression occurs after $n$ if and only if all lineages initiated by type-$E$ individuals up to that time go extinct.

Using $i = 1$ in (2) with the definitions of the life-history parameters in section 2, we arrive at the following expression:

$$\begin{aligned} F_{I_1(k,n)}(s_{k+1}, s_{k+2}, ..., s_n) =& (1-r_{1,k})(1-p_{1,k}) + (1-r_{1,k})p_{1,t}F_{I_1(k+1,n)}(s_{k+2}, ..., s_n) \\ &+ r_{1,t}G_1(k; p_{0,t}s_{k+1} + 1 - p_{0,t}). \end{aligned} \tag{4}$$

This can be computed for all $k < n$ using $F_{I_1(n,n)} = 1$ for any given sequence of environmental states.

It can be shown (see A.1) that the following holds.

$$F_{I_0(0,n)}(s_1, s_2, ..., s_n) = \prod_{l=0}^{n-1} G_0(l; p_{0,l}F_{I_1(l,n)}(s_{l+1}, s_{l+2}, ..., s_n) + 1 - p_{0,l}) \tag{5}$$

which can be computed using (4) and a sequence of environmental states.

The hazard rate conditioned on the environment, $\tilde{H}_n$ (with $n \in \mathbb{N}_0$), is the probability that introgression occurs at a time $n$, given that it has not occurred before, and given the environmental sequence. Using this definition with (3) and (5), the following expression is reached:

$$\tilde{H}_n = 1 - \frac{\prod_{l=0}^{n-1} G_0(l; p_{0,l}F_{I_1(l,n)}(s_{l+1}, s_{l+2}, ..., s_n) + 1 - p_{0,l})}{\prod_{l=0}^{n-2} G_0(l; p_{0,l}F_{I_1(l,n-1)}(s_{l+1}, s_{l+2}, ..., s_n) + 1 - p_{0,l})} \tag{6}$$

Note that $\tilde{H}_n$ is a random variable. The hazard rate can be simulated to study its distribution, by using the procedure outlined above, which we summarize here:

(1) Simulate an environmental sequence.
(2) Use this environmental sequence to calculate (1) backwards in time. It is important to choose a suitably large value of time from which to start so that (1) sufficiently converges to the asymptotic extinction probability.
(3) Use the values of (1) to calculate (4) for all values $k$ and $n$.
(4) Use the results from calculating of (4) to calculate the hazard rate in (6)

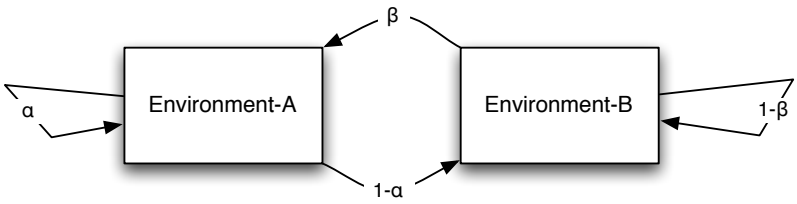


FIGURE 2. A schematic representation of the environment states and their transition probabilities.

(5) Store this value of the hazard rate and loop from step 1 again to arrive and multiple realizations of (6) from which expectations may be taken.

## 4. RESULTS

To illustrate the methodology, we will consider a situation where the environment changes according to a two-state Markov chain, with states $A$ and $B$. The transition probability from state-$A$ to itself is written as $\alpha$, and the transition probability from state-$B$ to state-$A$ as $\beta$. The scheme is summarized in Fig. 2. The initial environment at time zero is chosen from the stationary distribution of environmental states:

$$\pi_A = \frac{\beta}{1-\alpha+\beta}, \quad \pi_B = \frac{1-\alpha}{1-\alpha+\beta}, \tag{7}$$

and then evolves according to the defined transition probabilities. We now investigate how environmental randomness and plant life-histories together affect introgression risks. We assume that environment-A is a favorable environment and environment-B an unfavorable environment, so germination rates, hybridization rates, survival probabilities, and number of seeds produced by different plants are larger when the environmental state is A.

For convenience, all offspring distributions in the model are taken to be Poisson. The p.g.f. of a Poisson-distributed random variable is given by:

$$G(s) = e^{-m(1-s)}, \tag{8}$$

where $m$ denotes the expectation. The means of $\xi_{i,j}$ ($i \in \{0, 1, E\}$, $j \in \{A, B\}$) are denoted by $m_{i,j}$.

Figure 3(a) shows how the hazard rate changes with time. The hazard rate averaged over all environments is zero for the first year, because it is impossible to create a type-$E$ individual after just one year (as shown in Fig. 1), but then quickly approaches an asymptote. However, the dynamics for a specific environmental sequence can be much more capricious, and this leads to a large variance as is also shown in Fig. 3(a).

### 4.1. **Hazard rates in random and deterministically varying environments.**

For comparison, Fig.3(b) shows the hazard rates in deterministically alternating environments when the starting condition is state A (blue) or B (red). As can be seen, the average level is nearly the same, but the deviations from the mean

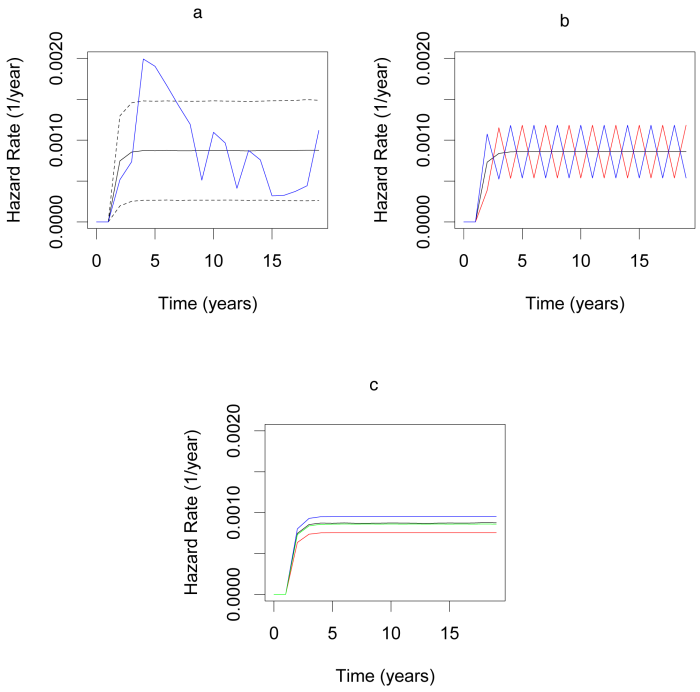


FIGURE 3. The hazard rate for different environmental processes. (a) Random environments, with a realization (blue), the mean (black) and standard deviations above and below the mean (dotted). (b) Deterministically alternating environments, starting with environmental state A(blue) or B(red), and the time-average (black). (c) Mean hazard rate in the case of random environments (black), deterministically alternating environments (green), and time-homogeneous environments with life-history parameters equal to the arithmetic mean (blue) and the geometric mean (red). Parameter values: $\alpha = \beta = 0.5$, $m_{0,A} = 10$, $m_{0,B} = 5$, $m_{1,A} = 1100$, $m_{1,B} = 800$, $m_{E,A} = 1400$, $m_{E,B} = 1000$, $p_{1,A} = p_{1,B} = p_{E,B} = 0.8$, $p_{E,A} = 0.9$, $r_{1,A} = r_{1,B} = r_{E,A} = r_{E,B} = 0.8$

in the random environment situation are much higher than those in the alternating environment case. Thus, random environments can induce periods of much higher risk. In Fig.3(c) we show the average hazard rates for respectively random environments, alternating environments, and constant environments with life history parameters equal to the geometric time-average, and to the arithmetic time-average. The average hazard rates are similar for random and alternating environments. Arithmetic time-averaged environments give a higher-than-average hazard rate, and geometric time-averaged environments a smaller one.

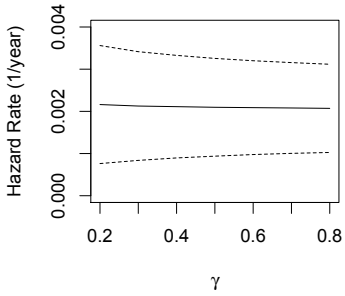


FIGURE 4. Effect of environmental autocorrelation on the hazard rate. Parameter values: $m_{0,A} = 10$, $m_{0,B} = 5$, $m_{1,A} = 1200$, $m_{1,B} = 800$, $m_{E,A} = 1400$, $m_{E,B} = 1200$, $p_{0,A} = 0.001$, $p_{0,B} = 0.0009$, $p_{1,A} = 0.9$, $p_{1,B} = 0.8$, $r_{1,A} = r_{e,A} = r_{e,B} = 0.8$, $r_{1,B} = 0.7$. The solid black line represents the mean hazard rate. The dotted lines represent one standard deviation over and under the average hazard rate.

4.2. **Environmental autocorrelation.** To examine the effects of environmental autocorrelation on the hazard rate, we take $1 - \alpha = \beta = \gamma$, with $\gamma \in (0, 1)$. Under this scheme, the expected amount of time in each environment state is the same. If $\gamma = 0.5$, the environment states are independent. They are negatively autocorrelated if $\gamma > 0.5$, and positively autocorrelated if $\gamma < 0.5$. Fig. 4 shows the effects of $\gamma$ on the hazard rate. As illustrated, the autocorrelation does not affect the mean hazard rate very much, whereas increasing $\gamma$ causes a small decrease in its standard deviation. This implies that for the parameter combinations that we investigated overall introgression risk is reduced in autocorrelated environments.

4.3. **Life history parameters.** The effect of life-history parameters depends on the environmental process. To illustrate this, we consider the effect of flowering probability $r_{1,B}$ and survival probability $p_{1,B}$ on the hazard rate Varying other flowering and survival probabilities leads to similar effects.

Figure 5 shows the hazard rate as a function of the flowering probability $r_{1,B}$ for different environmental scenarios and different values of $p_{1,B}$. As seen in the figure, the effect of flowering probability on the hazard rate depends on the combination of $\beta$ and $p_{1,B}$. When the survival probability in bad environments is low ($p_{1,B} = 0.1$, Figs.5 (a) and (c)) the hazard rate increases with increasing flowering probability, regardless of the value of $\beta$.

Figure 5 (b), on the other hand, shows an scenario where, given that one is in an unfavorable environment, one is likely to stay in that environment ($\beta = 0.1$), and also the survival probability of a non-flowering plant is high (0.9). In this scenario, the hazard rate is more or less independent of $r_{1,B}$.

As shown in Fig. 5 (d), the asymptotic level of the mean hazard rate can also decrease with $r_{1,B}$. This happens when the probability that an environment

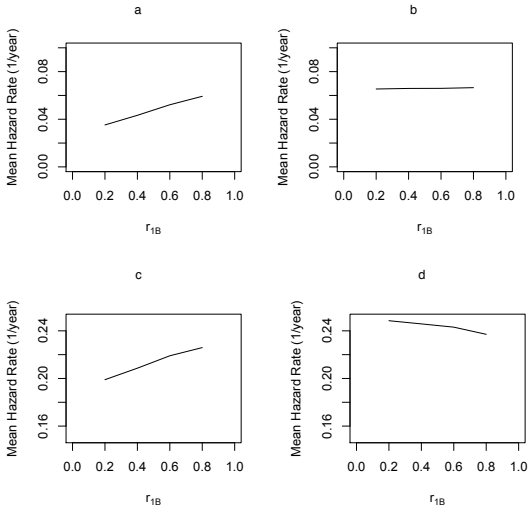


FIGURE 5. Effects of flowering probabilities $r_{1,B}$ on the mean hazard rate. Parameter values: $\alpha = 0.5$, $m_{0,A} = 1400$, $m_{0,B} = 1000$, $m_{1,A} = 1200$, $m_{1,B} = 550$, $m_{E,A} = 1300$, $m_{E,B} = 1200$, $p_{0,A} = 0.001$, $p_{0,B} = 0.0009$, $p_{1,A} = 0.9$, $p_{E,A} = 0.9$, $p_{E,B} = 0.8$, $r_{1,A} = 0.8$, $r_{E,A} = r_{E,B} = 0.8$, with subplot (a) showing $\beta = 0.1$, $p_{1,B} = 0.1$, subplot (b) showing $\beta = 0.1$, $p_{1,B} = 0.9$, subplot (c) showing $\beta = 0.9$, $p_{1,B} = 0.1$ and subplot (d) showing $\beta = 0.9$, $p_{1,B} = 0.9$

changes from unfavorable to favorable is high, and the probability of surviving is also high.

## 5. Discussion

In this paper we extended the methodology first presented in Ghosh and Haccou (2010) to deal with repeated invasions to situations with environmental stochasticity. This has led to many new, and sometimes surprising, results concerning introgression risk. As shown in subsection 4.1, introgression risks in random environments can be very different from those in deterministically varying environments, or in constant deterministic environments with the same time-averaged values of the life history parameters (see Fig. 3). As can be expected, the average hazard rate is higher in constant environments with the arithmetic mean life history parameters, since extinction probabilities of invasions are lower in such environments (see e.g. Haccou and Iwasa (1996)). Average hazard rates are lower in environments with geometric time-averaged parameters than in the random environment. Thus, using the arithmetic mean hazard rate overestimates the mean risk, whereas the geometric mean hazard rate underestimates it. As shown in Fig. 3(a), however, the hazard rate in random environments varies much in time, and there can be times at which it is much higher than its mean. Therefore, average

hazard rates are generally not a good measure of risk in random environments. We recommend that at least the variance of the hazard rate distribution is also taken into account. A closer examination of the distribution of the asymptotic hazard rate is also possible with the methods that we presented in this paper. In this way for instance a 95 percent upper bound can be established for the value of the hazard rate, which would provide a conservative estimate of the risk.

The value of the hazard rate is slightly affected by environmental autocorrelation (Fig. 4). We found that risks are reduced in strongly autocorrelated environments. This agrees with the results of Haccou and Vatutin (2003), who showed that success of sequential invasions is lower in environments with stronger positive autocorrelation, due to the increased length of 'runs of bad luck' in such environments. The effect that we found is, however, quite small.

Besides affecting the magnitude of introgression risks, random environments also change the effects that life history parameters have on introgression risks, as shown in subsection 4.3. If a plant in a poor environment can expect to be in a better environment where it would have a higher fecundity the next year, then it is better not to delay flowering. In this case, introgression riks decreases with flowering probability, as shown in Fig. 5 (d). However in delaying flowering, the plant also exposes itself to the risk of not surviving to the next flowering season. Thus, if survival probability is low, it is better to flower quickly. In this case introgression risk will increase with flowering probability (see Figs. 5 (a,c)). When the probability that a bad environment improves is low, and the survival probability of a non-flowering plant in a bad environment is high, it makes little difference if a hybrid flowers or not, and the hazard rate is more or less independent of $r_{1,B}$ (Fig. 5 (b)). Consequently, there is an interactive effect of environmental stochasticiy, flowering probabilities, survival probabilities and plant fecundities. As a result, the consequences of life history parameters for introgression risks are much more difficult to predict in random environments. This is a subject of further study.

While some aspects of environmental stochasticity were considered in introgression studies before by, for example, Davis et al. (1999) and Thompson et al. (2003), the combined effect of environmental and demographic stochasticity has not been examined before. In this paper we combine the two sources of stochasticity. Davis et al. (1999) found that environmental stochasticity increases the introgression rate. Our results suggest that this is not the whole story—while there might be periods where environmental stochasticity increases hazard rates, there can also be prolonged times of low introgression risk. Note, however, that the approach we present is very different to theirs. They consider the time for the wild-population to contain 90 percent of the transgene as a measure of introgression rate, whereas we consider the hazard rate of a permanent lineage being formed. In our view, the hazard rate is a less arbitrary measure, since it does not involve choosing a threshold frequency in the way that Davis et al. (1999) does. Even though Davis et al. (1999) and Thompson et al. (2003) fail to incorporate demographic stochasticity, they do include more specific information about the number of individuals carrying a transgene in a wild-population, whereas our approach concentrates on the risk of a transgene escaping in the first place. The growth of the transgene

frequency in the population given a successful invasion is another matter, that remains to be examined.

Even though this paper has outlined how to handle several types of stochasticity in introgression models, further research is needed to have a complete understanding of the mechanisms involved in introgression. For instance, potential invading genes will be linked to other genes, which can affect introgression risks. We are currently working on extending the methodology to incorporate such genetics, and to include effects of drift in small wild populations.

Another important generalization would be to incorporate multiple wild populations and to investigate how a metapopulation structure affects the spread of invading genes, which would involve elaborating on work by Hanski et al. (1999).

We would like to conclude by remarking that much must still be done to develop full models of introgression, but effects of demographic and environmental stochasticity are important aspects to include in introgression studies.

## 6. Acknowledgments

This research was funded through the research program 'Ecology Regarding Genetically modified Organisms (ERGO)', commissioned by four Dutch ministries. This funding program is managed by the Earth and Life Sciences Council (ALW) of the Netherlands Organisation for Scientific Research (NWO). P. Haccou's research is additionally supported by the NDNS (Nonlinear Dynamics of Natural Systems) program of NWO. M.C. Serra would like to thank the Fundação para a Ciência e Tecnologia for financial support through the scholarship SFRH/BPD/47615/2008.

## Appendix A. Appendix

A.1. **Derivation of** (5)**.** Putting $i = 0$ and $k = 0$ in (2) with the definitions from section 2, we find the following:

$$\begin{aligned}
&F_{I_0(0,n)}(s_1, s_2, ..., s_n) = F_{I_0(1,n)}(s_2, s_3, ..., s_n) G_0(0; p_{0,0} F_{I_1(1,n)}(s_2, s_3, ..., s_n) + 1 - p_{0,0}) \\
&= F_{I_0(2,n)}(s_3, s_4, ..., s_n) G_0(0; p_{0,0} F_{I_1(1,n)}(s_2, s_3, ..., s_n) + 1 - p_{0,0}) \\
&\qquad \times G_0(1; p_{0,1} F_{I_1(1,n)}(s_2, s_3, ..., s_n) + 1 - p_{0,1}) \\
&= \prod_{l=0}^{n-1} G_0(l; p_{0,l} F_{I_1(l,n}(s_{l+1}, s_{l+2}, ..., s_n) + 1 - p_{0,l}) \qquad (9)
\end{aligned}$$

## References

Davis, S.A., Catchpole, E.A., Pech, R.P., 1999. Models for the introgression of a transgene into a wild population within a stochastic environment, with applications to pest control. Ecol. Model. 119, 267-275

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. 301, 539-563.

Ghosh, A., Haccou, P., 2010. Quantifying stochastic introgression processes with hazard rates, Theor. Popul. Biol. 77, 171-180.

Ghosh, A., Serra, M.C., Haccou, P., In Press. Quantifying time-inhomogenous stochastic introgression risks with hazard rates, Theor. Popul. Biol.

Ghosh, A., Meirmans, P., Haccou, P., submitted. Quantifying introgression risk with realistic population genetics.

Haccou, P., Iwasa, Y., 1996. Establishment probability in fluctuating environments: a branching process model, Theor. Popul. Biol., 50, 254-280.

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

Haccou, P., Vatutin, V., 2003. Establishment success and extinction risk in autocorrelated environments, Theor. Popul. Biol., 64, 303-314.

Huxel, G.R., 1999. Rapid displacement of native species by invasive species: Effects of hybridization. Biol. Conservat. 89, 143-152.

Hails, R.S., Morley, K., 2005. Genes invading new populations: A risk assessment perspective. Trends Ecol. Evol. 20, 245-252.

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

Kalbfleisch, J.D., Prentice, R.L., 2002. The statistical analysis of failure time data, 2nd ed. John Wiley and Sons, New York.

Kwit, C., Moon, H.S., Warwick, S.I., Stewart Jr., C.N., 2011. Transgene introgression in crop relatives: molecular evidence and mitigation strategies, Trends Biotechnol., 29, 284-293.

Reichmann, J.R., Watrud, L.S., Lee, E.H., Burdick, C.A., et al., 2006. Establishment of transgenic herbicide-resistant creeiping bentgrass (*Agrostic stolonifera L.*) in nonagronomic habitats. Mol. Ecol. 15(13), 4243-4255.

Thompson, C.J., Thompson, B.J.P., Ades, P.K., Cousens, R., Garinier-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

## ENGLISH SUMMARY

The majority of the world's most important crops hybridise readily with their wild relatives. With the advent of genetically modified crops, the possible consequences of such hybridisation has come under increasing scientific scrutiny. One important possible consequence is introgression. Introgression is the permanent incorporation of genes from the genome of one population into another. Will genetically modified genes introgress from cultivated plants into their wild relatives? If so, when will this happen? What is a suitable measure for the risk for introgression? To answer these questions requires a combination of experimental and theoretical studies. Experimental approaches can help determine the fitness effects that a crop gene will have when placed in wild individuals, and theoretical approaches can then use this information to forecast the population growth of foreign genes in a wild population. The work in this thesis provides a suitable theoretical framework for quantifying the risk of occurrence of introgression, and companion projects present complementary experimental methods.

*Chapter 1* shows that if an invading gene has fitness benefit, then repeated hybridisation will result in the gene invading after some time. This is not to say that the risk of introgression of the gene is high—the gene will eventually invade, but it might take a long time to do so. The risk of introgression occurrence is governed by how long it takes for a permanent lineage to be formed. Since the number of invaders can be initially small, then randomness in the number of offspring (so-called demographic stochasticity) is a crucial factor in determining when permanent lineages form. The theory of branching processes is used to develop a methodology for calculating the time at which permanent invading lineages are formed. From this time distribution, a measure of introgression risk is introduced: the hazard rate. The hazard rate of introgression is the probability that introgression occurs at a certain time given that it has not occurred before. A sample calculation of the hazard rate is shown for plants which can either flower after one year or delay flowering. It is shown that introgression risks can sometimes be higher if plants delay flowering instead of flowering immediately.

*Chapter 2* uses the results of chapter 1 and combines them with the preliminary results from a companion study to estimate introgression risk from cultivated carrots into their wild relatives. A sensitivity analysis was performed to determine the most important factors driving introgression. The combination of the experiments and theory in this chapter hints at ongoing work, which uses the carrot (*Daucus carota*) as a model species in developing a comprehensive methodology to estimate introgression risks.

*Chapter 3* elaborates on the the results from chapter 1 by allowing for hybridisation rates to change in time, allowing for the incorporation of management strategies such as crop-rotation to be included into risk calculations. The case where it takes several generations before a fitness advantage is seen is also investigated.. Hazard rates of introgression can change in time during crop rotations,

so the task of choosing a suitable level of introgression risk is complicated. Procedures for averaging the hazard rate over time are presented. The average hazard rate could be used as a measure for introgression risk, but it might be misleading since it can significantly underestimate introgression risks during some time periods. Randomly varying hybridisation rates, due to chance changes in pollinator activity or shifts in weather, are also investigated.

*Chapter 4* presents a framework to incorporate genetics into the previous methodologies. Furthermore, procedures for calculating hazard rates using computer simulations are shown, in addition to the mathematical methods presented in previous chapters. In order to retain mathematical tractability when using branching processes, the work in previous chapters assumed that introgression was occurring into (infinitely) large wild populations. The use of simulation-based techniques allowed tests to be done on how large a wild population has to be before the assumptions from previous chapters hold. Both branching process and computational approaches give similar predictions for population sizes on the order of 100. For small wild populations, introgression is primarily driven by chance and is less dependent on fitness effects of invading genes. For large wild populations, introgression is driven by selection and branching processes are an efficient tool to calculate hazard rates. Genetic linkage is found to be an important factor affecting introgression risk.

*Chapter 5* generalises the approaches from chapters 1 and 3 to allow for random environments. In such scenarios, the hazard rate can change randomly in time. This means that there may be some periods where introgression risk may be higher than in others. This leads to a practical challenge in choosing an acceptable level of introgression risk: should one choose an average hazard rate, a maximum hazard rate or some other level? If an average hazard rate is used, introgression risk might be severely underestimated for some periods of time. Increasing flowering probabilities can either increase or decrease the average hazard rate, depending on the environment.

# SAMENVATTING

Het merendeel van de belangrijkste cultuurgewassen in de wereld hybridiseren gemakkelijk met hun wilde verwanten. Met de komst van genetisch gemodificeerde gewassen, is het belangrijk om te onderzoeken wat de mogelijke gevolgen van dergelijke hybridisatie zijn. Een belangrijk mogelijk gevolg is introgressie. Introgressie is de permanente incorporatie van genen van het genoom van een bepaalde populatie in een andere populatie. Zullen genetisch gemodificeerde genen van cultuurgewassen op den duur permanent in hun wilde verwanten voorkomen? Zo ja, wanneer zal dit gebeuren? Wat is een geschikte maat voor het risico op introgressie? Om deze vragen te beantwoorden is een combinatie van experimenteel en theoretisch onderzoek nodig. Experimentele benaderingen zijn nodig om de fitnesseffecten te bepalen, die een bepaald gen zal hebben indien het in het genoom van een wilde verwant terecht komt. Deze informatie kan dan vervolgens in theoretisch onderzoek worden gebruikt om het verloop van de frequentie van zo'n gen in een natuurlijke populatie te voorspellen. Het onderzoek dat in dit proefschrift wordt beschreven biedt een geschikt theoretisch kader voor het kwantificeren van het risico op het optreden van introgressie. Dit onderzoek maakt deel uit van een groter programma, met complementaire empirische deelprojecten.

In *Hoofdstuk 1* wordt uitgelegd dat wanneer een gen een fitness voordeel heeft, hoe klein ook, herhaaldelijke hybridisatie uiteindelijk zal resulteren in introgressie. Dit betekent echter niet noodzakelijk dat het risico op introgressie van het gen ook hoog is. Het kan namelijk ook heel lang duren voordat dit gebeurt. Het risico dat introgressie optreedt wordt bepaald door de duur van de periode voordat er een lijn van nakomelingen wordt geproduceerd waarin het gen blijft voorkomen . Dit noemen we een 'permanente lijn'. Aangezien het aantal hybrides aanvankelijk erg klein is, is inter-individuele variatie in overlevingskans en nakomelingen (zogenaamde demografische stochasticiteit) een essentiële factor bij dit proces. De theorie van vertakking processen wordt gebruikt om een methode af te leiden waarmee de verdeling van de tijd tot initiatie van een permanente lijn kan worden bepaald. De hazard rate van introgressie is de kans per tijdseenheid dat er een permanente lijn ontstaat, gegeven dat dat nog niet eerder is gebuerd. In dit hoofdstuk wordt een voorbeeld van een berekening van de hazard rate gegeven, voor planten die eenmalig bloeien, na een of meerdere jaren. Er wordt aangetoond dat het uitstellen van de bloei het risico van introgressie kan verhogen.

*Hoofdstuk 2* maakt gebruik van de resultaten van hoofdstuk 1 en combineert ze met de voorlopige resultaten van een empirisch onderzoek naar introgressie van genen van gecultiveerde peen in hun wilde verwanten. Een gevoeligheidsanalyse werd uitgevoerd om de belangrijkste factoren die introgressie beinvloeden vast te stellen. De combinatie van de experimenten en de theorie in dit hoofdstuk verwijst naar lopend onderzoek, waarin de wortel wordt gebruikt (*Daucus carota*) als modelsoort in de ontwikkeling van een uitgebreide methodologie om introgressie risico's te bepalen.

*Hoofdstuk 3* worden de resultaten van hoofdstuk 1 gegeneraliseert, door toe te staan dat de verdeling van het aantal gevormde hybriden kan veranderen in de tijd. Hierdoor kunnen effecten van gewas management, zoals gewas-rotatie op introgressie risico worden bepaald. Hazard rates van introgressie kunnen hierdoor variëren in de tijd. Dit compliceert het bepalen van een geschikt niveau van introgressie risico. In dit hoofdstuk wordt een procedure om het gemiddelde van de hazard rate te berekenen gepresenteerd. Dit zou kunnen worden gebruikt als een maat voor introgressie risico, maar dat zou misleidend kunnen zijn, aangezien er periodes zijn waarin het risico aanzienlijk hoger is. Het effect van toevalsfluctuaties in hybridisatie-snelheden, bijvoorbeeld ten gevolge van variatie in de activiteit van bestuivers, en/ of weersomstandigheden, worden ook onderzocht. In dit hoofdstuk worden ook situaties onderzocht waarin het enkele generaties duurt voordat er een fitness voordeel is.

*Hoofdstuk 4* biedt een framework om meer gecompliceerde genetische mechanismen te beschouwen. Bovendien worden procedures voor de berekening van de hazard rate op grond van computersimulaties gepresenteerd, naast de wiskundige methoden van de voorgaande hoofdstukken. In de vorige hoofdstukken werd ervan uit gegaan dat de wilde populatie groot genoeg was om de kans op interactie tussen hybride indivduen te verwaarlozen. Met behulp van simulatie-gebaseerde technieken kan worden gedaan hoe groot een wilde populatie moet zijn voordat deze aanname een redelijke benadering geeft. Het blijkt dat de methode gebaseerd op vertakkingsprocessen en de computersimulaties vergelijkbare uitkomsten geven zodra de populatieomvang van de orde van grootte van 100 individuen of meer is. Voor kleine wilde populaties wordt introgressie in de eerste plaats gedreven door toeval en is het minder afhankelijk van fitness-effecten van de invasie van genen. Voor grote wilde populaties, wordt introgressie gedreven door selectie, en zijn vertakking processen een doeltreffend instrument om risico's te berekenen. Genetic linkage tussen loci van cultuurgewassen blijkt een belangrijke factor invloed op introgressie risico te hebben. Dat betekent dat koppeling tussen een gemodificeerd gen en een gen dat in de natuur nadelig is een bruikbare strategie kan zijn om introgressie risico te verlagen.

*Hoofdstuk 5* generaliseert de benaderingen van de hoofdstukken 1 en 3 naar situaties met omgevingen met toevalsfluctuaties. In dergelijke scenario's kan de hazard rate willekeurig veranderen in de tijd. Het blijkt dat er perioden kunnen zijn waarin introgressierisico's behoorlijk hoger zijn dan gemiddeld. Dit leidt tot een probleem bij het kiezen van een aanvaardbaar niveau van introgressie risico: moet men kiezen voor een gemiddelde hazard rate, een maximale hazard rate of een ander niveau? Indien de gemiddelde hazard rate wordt gebruikt, kan introgressie risico ernstig worden onderschat enige tijd. Ook blijkt dat het effect van life history parameters, zoals de kans op bloei, afhangt van de manier waarop de omgeving varieert in de tijd.

## ACKNOWLEDGMENTS

Several people were instrumental in bringing this work to its current state. Patsy Haccou's contribution to every chapter should not be understated. A collaboration with Patrick Meirmans brought chapter 4 into fruition, and chapters 3 and 5 were very much a team effort involving São Serra.

Tom de Jong, Wil Tamis, Klaas Vrieling, Jun Rong and Cilia Grebenstein provided helpful guidance throughout my project.

The work presented would not have been possible without funding from the Ecology Regarding Genetically Modified Organisms (ERGO) program, which is managed by the Earth and Life Sciences Council (ALW) of the Netherlands Organisation for Scientific Research (NWO).

## CURRICULUM VITAE

Atiyo Ghosh was born on 24 September 1984 in Toronto, Canada. In 2003, after completing his A-Levels at Harrow School, London, he pursued Bachelor's and Master's degrees in Physics at the University of Cambridge. During these degrees, his interests in Biology and stochasticity were piqued by courses in Evolution and Behaviour and Quantum Physics. In 2008, after completing his studies at Cambridge, he moved to the Institutes of Biology and Environmental Sciences at Leiden University to start the work contained within this thesis. He is planning to continue with post-doctoral research on stochastic modelling with biological applications in Japan.