Taylor & Francis
Taylor & Francis Group

Letters in Biomathematics, 2015
Vol. 2, No. 1, 13–28, http://dx.doi.org/10.1080/23737867.2015.1040859

RESEARCH ARTICLE

Modelling the evolution and expected lifetime of a deme's principal gene sequence

B.K. Clark*

Department of Physics, Illinois State University, Normal, IL, USA

(*Received 3 April 2015; accepted 10 April 2015*)

The principal gene sequence (PGS), defined as the most common gene sequence in a deme, is replaced over time because new gene sequences are created and compete with the current PGS, and a small fraction become PGSs. We have developed a set of coupled difference equations to represent an ensemble of demes, in which new gene sequences are introduced via chromosomal inversions. The set of equations used to calculate the expected lifetime of an existing PGS include inversion size and rate, recombination rate and deme size. Inversion rate and deme size effects are highlighted in this work. Our results compare favourably with a cellular automaton-based representation of a deme.

Keywords: inversion; recombination; gene sequence; stasipatric speciation

1. Introduction

Chromosomal inversions are examples of a broader group of chromosomal rearrangements. Stasipatric speciation is when chromosomal inversions are the primary mechanism of speciation (White, 1978); however, examples of speciation typically include chromosomal inversions only as a contributing factor. Examples are ubiquitous, including Drosophila (Ayala & Coluzzi, 2005; Noor, Grams, Bertucci, & Reiland, 2001; Ranz et al., 2007), sunflowers (Rieseberg, 2001; Rieseberg et al., 2003), and primates (Ayala & Coluzzi, 2005; Lee, Han, Meyer, Kim, & Batzer, 2008; Navarro & Barton, 2003a, 2003b), for example. Suppressed recombination of inverted chromosome segments has been investigated as a speciation mechanism (Ayala & Coluzzi, 2005; Navarro & Barton, 2003a, 2003b), and the precise role of inversions in speciation is still debated (Faria & Navarro, 2010). In part this is because the role of chromosomal inversions in speciation is difficult to quantify in nature. For example, its role is discussed qualitatively in conjunction with the hybrid dysfunction model of speciation (Ayala & Coluzzi, 2005).

In a simplified model of a reproducing population of a species or deme, a chromosomal rearrangement can enter the deme as a chromosomal inversion. In this example, the chromosomal inversion competes with all other chromosomal gene sequences until all other sequences are replaced. At this point the new chromosomal gene sequence is said to be fixed. An early model of chromosomal rearrangement fixation in a deme was provided by Lande (1979), when he extended work on the fixation of new gene mutations (Kimura, 1962) to chromosomal rearrangements. Kimura's model (1962) included a diffusion approximation based on the Wright–Fisher (Fisher, 1922; Wright, 1931) model of genetic drift. Later, fixation probabilities for new chromosomal arrangements were investigated (Hedrick, 1981;

*Email: bkc@phy.ilstu.edu

Spirito, 1992, 1998) and showed that the calculated fixation probabilities for the early models were in qualitative agreement with each other. However, the results were not compared to biological systems and predated computational simulations of chromosomal rearrangements in a deme.

Chromosomal inversions in an isolated deme are estimated to occur with a rate of 10^{-4} to 10^{-3} (gamete gen)$^{-1}$ (Lande, 1979). This low rate has allowed models of inversion fixation to assume only one chromosomal inversion (or rearrangement of any kind) is present in a deme at a time. Inversions can also be introduced into demes via immigration at rates at least as great as 0.05 (gamete gen)$^{-1}$ (Beerli & Felsenstein, 1999; Hey & Nielsen, 2004; Wang & Whitlock, 2003). The fixation probability of a new inversion is usually treated as a function of deme size and inversion fitness (Hedrick, 1981, 1992; Spirito, 1998). Even when the inversion is otherwise fitness neutral, the size of the inversion directly contributes to a reproductive advantage of smaller inversions over larger inversions in a computer simulation Clark, Wabick, and Weidner (2012) and in Drosophila (York, Durrett, & Nielsen, 2007). This reproductive difference can also be treated as fitness. The statistical properties of the reproduction process also affect fixation probabilities. When applied to inversions, the Wright-Fisher model assumes reproductive competition between the original and new inversions is described by a binomial probability distribution, although this is one of several possible models (Der, Epstein, & Plotkin, 2011; Schweinsberg, 2003) of reproduction.

The historical focus of research on inversions has been the fixation probability of a single inversion (Hedrick, 1981; Lande, 1979; Spirito, 1992, 1998), the linkage between inversions and individual gene variants (Noor et al., 2001; Rieseberg, 2001), and the role of inversions in evolution (Kirkpatrick, 2010; Lowry & Willis, 2010). Additionally, most of these models (Lande, 1979; Spirito, 1992, 1998) calculated endpoints but not the trajectories taken by a deme or ensemble of demes to reach the endpoints. Hedrick (1981) did look at a limited number of trajectories for an infinite population size in which stochastic fluctuations can be ignored.

We model the trajectory of the principal gene sequence (PGS), defined as the gene sequence with the highest occurrence frequency in the present generation, so that we may better understand the transition of a gene sequence to and from the PGS state. There are many gene sequences within a deme that fail to survive to the point of fixation. Consequently, it is reasonable to look at the process as a fixed gene sequence, the PGS, is replaced by one of many new gene sequences that appear in the deme over many generations. To accomplish this task, a computational model of a single deme in which inversion is allowed is implemented. This simulation allows us to follow the evolution of the PGSs in the deme over many generations. In particular, we compute the lifetime of each PGS in the deme as a function of inversion rate and deme size. Individual trajectories can be interesting, but provide an incomplete picture. We introduce a set of coupled equations that approximate the computer simulation. The solution to the set of coupled equations yields the density of PGS states for the ensemble and the PGS lifetime. Here, a PGS state is defined by the fraction of strands with the PGS in the deme and the density of PGS states is the state distribution of non-interacting demes in the ensemble. Simulations include genetic drift, limited fitness effects and multiple inversions.

2. Model

2.1. Cellular automaton

A cellular automaton (CA) model begins by establishing a grid of cells. Here, each individual in the deme is represented by a row of cells and each cell in row represents a specific gene.

Table 1. Definitions of abbreviation, parameters, and variables.

B_{ki}	Probability of deme with size n and i PGS strands producing a deme of the same size with k PGS strands in the next generation
C_{kn}	Ratio of $\rho_k(t)/\rho_n(t)$
CA	Cellular automaton model
DGS	Different genetic sequences
ϕ	The absolute value of the fractional part of Γ
Γ	Expected change in the size of the subpopulation with the PGS
I	Chromosomal inversion rate set between 0.0001 (strand gen)$^{-1}$ and 0.05 (strand gen)$^{-1}$ for this work
j	Number of strands with the PGS
j_{eq}	Equilibrium value of j
K	Inverse lifetime of PGS
L	Loss rate in recombination calculated to be 0.140 (pair gen)$^{-1}$ for this work
m	Number of genes per chromosome set to 50 for this work
n	Number of haploid strands in a deme varies between 20 and 80 in this work
q	Minimum number of strands for gene sequence to be a PGS
PGS	Principal gene sequence
$\rho_k(t)$	Fraction of the ensemble of demes with k PGS members
$\Delta\rho_k(t)$	Change in the value of $\rho_k(t)$ in one generation
R	Recombination rate during reproduction set to 1 (pair gen)$^{-1}$ for this work
SGS	Same genetic sequence
σ_{ki}	Fraction of $\rho_i(t)$ transferred to $\rho_k(t)$ in one generation
τ	Lifetime of a PGS
U	Fixation probability of underdominant chromosomal inversion
ω	Absolute value of the whole number part of Γ

The cell sequence in the CA is the same as the gene sequence for the individual. The major features of the model used in this work have been previously discussed in detail Clark et al. (2012). Table 1 summarizes the abbreviations, parameters and variables used in this work. Each member of a deme of n haploid individuals is modelled as a single strand of DNA divided into m genes. The simulations in this work are completed for $20 \leq n \leq 80$ and $m = 50$. Initially, the gene sequence is identical for all n strands and the initial position of each gene serves as the gene trait assignment. A strand must have one gene from each of the m traits to survive and reproduce. As time advances individual genes move from their original positions, but retain their original trait assignments. Repeating cycles of inversion and recombination lead to the replacement of the PGS with a new sequence that differs from the existing PGS only in gene order.

The reproduction process begins by randomly selecting two strands with complete replacement to reproduce. All viable strands are equally likely to be selected for reproduction and will be selected slightly more than once to account for the nonviable strands. The recombination rate R is 1 (pair gen)$^{-1}$ in this work, and the recombination locus is picked at random from a uniform distribution of the set of m loci consisting of all loci between adjacent genes and the strand end at gene position 1. Two offsprings are formed from each recombination and saved as members of the next generation. The reproduction process continues until the new generation consists of n strands. Offsprings that lack a complete set of genes are assigned as a reproduction probability of zero and are removed from the deme at the start of the next reproduction cycle. The parents are returned to the current generation

pool, where they may be selected to reproduce again. If a specific strand is selected to be a parent more than once, it will usually be paired with a different strand each time since both strands are randomly selected to reproduce. Each of the n strands in the new generation undergoes inversion with an average inversion rate I, where 0.0001 (strand gen)$^{-1} \leq I \leq$ 0.05 (strand gen)$^{-1}$ in this work. Two inversion points are randomly selected from a uniform distribution of the $m + 1$ loci between genes or at the ends of the strands for each pair of strands as the final reproduction step.

Figure 1(a) shows a section of a typical cellular automaton trajectory (top curve) for $I = 0.01$ (strand gen)$^{-1}$ and $n = 20$, where the number of strands with the PGS is presented as a function of time. A particular genetic sequence is usually the PGS for an extended period of time, but the number of strands with the PGS fluctuates. If the fluctuation is great enough that the number of strands with the PGS drops below $n/2$ a new PGS usually replaces the existing PGS. The tick marks at the bottom of Figure 1(a) denote the times when the existing PGS was replaced. Several genetic sequences may exist in the deme at the same time and compete for the role of PGS. In some cases, a PGS survives for a long period of time and in other cases a new PGS is rapidly replaced. A new genetic sequence is introduced at an average rate of one every five generations for this inversion rate and one-third of new inversions are expected to disappear from the deme in a generation.

An alternative view is presented in Figure 1(b), which shows the fraction of PGSs with a measured lifetime that equals or exceeds the indicated number of generations. The large initial decrease in PGS fraction reflects the competition between different genetic sequences for the role of PGS, when the PGS state is near $n/2$. When a genetic sequence first becomes a PGS, the displaced PGS is usually present with a high enough frequency to again become the PGS. After 20–40 generations one sequence usually has a much greater frequency than any other sequence so it becomes an established PGS and the magnitude of the slope of the curve in Figure 1(b) decreases and becomes constant. Consequently, genetic drift via inversion in this simulation occurs on two time scales. On a short time scale, a deme has multiple gene sequences and no one sequence has been established as the long term PGS; the system is called polymorphic. On a much longer time scale, a PGS exists for many generations in spite of competition with numerous new genetic sequences created via inversions. These inversions can be classified as rare, since they appear in only a small fraction of the deme membership and quickly disappear from the deme. Eventually, the frequency of some competing genetic sequence will exceed the existing PGS and it becomes the new PGS. The constant slope K corresponds to an exponential decay process. The lifetime, $\tau = -1/K$, is the time required for the fraction of original PGSs in the ensemble to decrease by a factor of $1/e$, when measured in the long time limit after the exponential tail has formed. For example, the exponential tail is seen to dominate the curve in Figure 1(b) after 25 generation and has $\tau = 259$. gen for $K = -0.00387$/gen, when measured for demes with PGSs that survive for between 200 and 1000 gen. Figure 1(b) only shows the first 100 generations to highlight the formation of the exponential tail.

2.2. Model approximations

Each step in the production of a new generation, including the occurrence of inversions, recombinations and pair selection for reproduction is stochastic. The associated stochastic fluctuations drive genetic drift. A set of equations describing the time evolution of a deme's PGS can be separated into terms that depend on the mean effects of inversion, recombination and pair selection and terms that incorporate the stochastic fluctuations present in the system. As the size of the deme becomes large, the importance of stochastic fluctuations is minimized

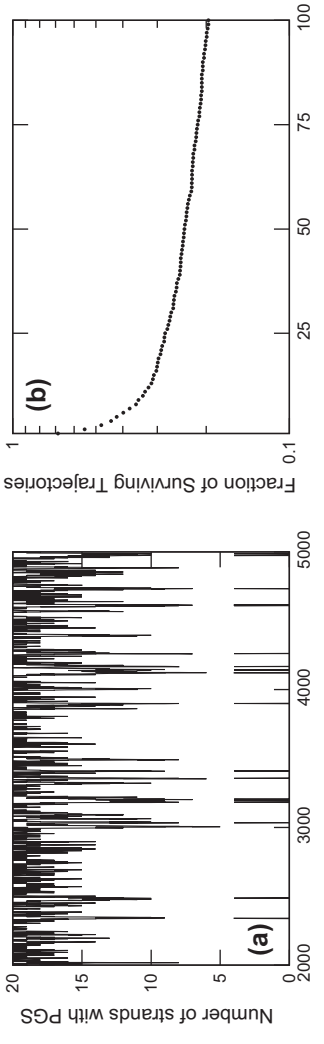

Figure 1. Typical results of cellular automaton model. A section of a sample trajectory is shown in (a). The tick marks at the bottom of (a) indicate when a PGS is replaced. The curve in (b) shows the fraction of surviving PGSs as a function of time for the complete trajectory in (a).

and differential equations based on average rates are expected to be valid. For small demes, on the order of $n = 20$, stochastic fluctuations can dominate average changes.

The deme can be modelled as two subpopulations, one includes j strands with the PGS and the other includes $n - j$ strands with different sequences. We consider two simplifying approximations, either all $n - j$ strands have the same genetic sequence (SGS) or they all have different genetic sequences (DGS). We develop the SGS model here, and show the changes required for the DGS model in Appendix 1. All $n - j$ strands are assumed to have the same genetic sequence when inversions occur infrequently, since selection removes most new inversion from the deme before the next inversion occurs. Inversion with rate I will reduce the j number of strands with the PGS by an amount, Ij. There are no inversion losses from the $n - j$ strands in the SGS model. This is a valid assumption when $j = n - 1$, because the inversion makes no difference in the model. Since this approximation is not true when $j \approx n/2$, the stability of the PGS is underestimated in generations when the deme is polymorphic as described in conjunction with Figure 1(b).

The probability of producing nonviable offspring during recombination is equal to the fractional size of the sequence mismatch between two gene sequences. The average rate of producing nonviable offspring from parental genetic sequences selected with a uniform distribution of inversion loci is $L = 0.140$ (pair gen)$^{-1}$ for $R = 1$ (pair gen)$^{-1}$, as shown in Appendix 2. This effective population loss includes $Lj(1 - j/n)$ from each of the two subpopulations because of recombination between the subpopulations. The same absolute number of strands is lost in recombination from both subpopulations, but a greater fraction of the $n - j$ strands is lost. This confers an effective fitness advantage to the PGS subpopulation over the subpopulation with the inverted gene sequence since the PGS subpopulation always accounts for at least half the deme size. The total loss in population from recombination is

$$n_{loss} = 2Lj\left(1 - \frac{j}{n}\right). \tag{1}$$

Recombination losses to the deme are replaced by allowing viable individuals to reproduce until the population is returned to n.

The expected change Γ in the size of the subpopulation with the PGS is

$$\Gamma = -Ij - Lj\left(1 - \frac{j}{n}\right) + \left(2Lj\left(1 - \frac{j}{n}\right)\right)\frac{\left(j - Ij - Lj\left(1 - \frac{j}{n}\right)\right)}{n - 2Lj\left(1 - \frac{j}{n}\right)}, \tag{2}$$

where the first two terms represent the inversion and recombination reductions in the number of strands with the PGS, respectively. The third term accounts for the proportional replacement of strands lost from the entire deme. It is the product of the number of strands lost from the deme and the proportion of strands with the PGS after the inversion and recombination events described by the first two terms in the equation. Setting $\Gamma = 0$ to obtain the steady state solution yields a quadratic equation whose stable equilibrium solution is

$$j_{eq} = \frac{n}{2}\left(\frac{3}{2} + \frac{1}{2}\sqrt{1 - \frac{8I}{L}}\right) \tag{3}$$

for $I \leq L/8$. If $I \ll L/8$, Equation (3) can be approximated as

$$j_{eq} = n\left(1 - \frac{I}{L}\right). \tag{4}$$

As the inversion rate is increased, the fraction of population with the PGS decreases. Similarly, a decreasing loss rate from recombination results in a decrease in PGS stability. Values of j greater or lesser than the solution to Equation (3) have Γ values that are negative or positive, respectively.

Equation (3) in the SGS model becomes

$$j_{eq} = \frac{n}{2}\left(\left(1+\frac{1}{n}\right)+\sqrt{\left(1-\frac{1}{n}\right)-\frac{4I}{L}}\right) \tag{5}$$

in the DGS model (see Appendix 1), and Equation (4) becomes

$$j_{eq} = n\left(1-\frac{I}{L\left(1-\frac{1}{n}\right)}\right) \tag{6}$$

when $I \ll L/4$.

The ratio I/L becomes the important parameter in Equation (3) through Equation (6). Since L is a function of R, it is necessary to only vary I to gain a basic understanding of the population dynamics.

2.3. Ensemble picture

A more complete understanding of the life history of a PGS is gained by considering an ensemble of initially identical non-interacting demes with n strands. At all times after the first generation the ensemble will consist of $n+1$ different $\rho_k(t)$, where $\rho_k(t)$ is the density of demes in state k and $0 \le k \le n$ is the number of strands with the original PGS. Demes with k near $n/2$ are in polymorphic states and demes with k near n are in rare states. The change in each $\rho_k(t)$ during one generation is

$$\begin{aligned}\Delta\rho_k(t) = &\sum_{i=0}^{n}\rho_i(t)\sigma_{ki} - \rho_k(t)\sum_{i=0,i\neq k}^{n}\binom{n}{i}\left(\frac{k}{n}\right)^i\left(1-\frac{k}{n}\right)^{(n-i)}\\ &+\sum_{i=0,i\neq k}^{n}\rho_i(t)\binom{n}{k}\left(\frac{i}{n}\right)^k\left(1-\frac{i}{n}\right)^{(n-k)}.\end{aligned} \tag{7}$$

The second term in Equation (7) is the amount of $\rho_k(t)$ that is transferred to all other $\rho_i(t)$ and the third term is the amount of all other $\rho_i(t)$ that is transferred to $\rho_k(t)$ because of stochastic effects. A binomial probability distribution is used because it describes selection with complete replacement, in agreement with the selection processes in the cellular automaton simulation. Of course the specific trajectory of any single PGS cannot be predicted; however, the trajectory of the ensemble can be. A member of the ensemble in state $k < n$ may end up in any PGS state with $k \le n$ because of the stochastic character of the frequency of strand selection for inversion and recombination and pairing of strands for recombination.

The first term in Equation (7) describes the mean effects of inversion, recombination and deme size. Here, σ_{ki} is the fraction of $\rho_i(t)$ transferred to $\rho_k(t)$ according to Equation (2) or Equation (A2) for the SGS or DGS models, respectively. σ_{kk} is negative and represents the fraction of $\rho_k(t)$ that is transferred to all other $\rho_i(t)$. The absolute value of Γ can be greater or less than one, depending on the choice of parameters. It is convenient to denote the absolute value of the whole number part of Γ as ω and the absolute value of the fractional

part as ϕ. Then $1-\phi$ is the fraction of $\rho_k(t)$ that is transferred to $\rho_{k\pm\omega}(t)$ and the remainder, ϕ, is transferred to $\rho_{k\pm(\omega+1)}(t)$ in this model. The sign agrees with the sign of Γ. In the limit of small deme sizes and inversion rates ω is normally zero, so the $1-\phi$ fraction of $\rho_k(t)$ remains in the same k state and the remainder is transferred to one of the $k\pm1$ states, in agreement with the sign of Γ. The set of $\rho_k(t)$ can be solved for each generation using any matrix package. SCILAB was used for this work.

Once the same genetic sequence is common to less than half of the population, it may no longer be the PGS. This is determined by direct observation in a cellular automaton trajectory. In the ensemble picture, the minimum number of strands $q \leq n/2$ must be specified to define when a PGS is replaced. Any member of the ensemble with fewer than q strands is deemed to have the original PGS replaced and is removed from the ensemble. The replaced quantity of PGSs in the ensemble is $\sum_{k=0}^{q-1}\rho_k(t)$. In the ensemble simulation, these $\rho_k(t)$ are set to zero for all time after $t=0$. In the SGS model the PGS has been replaced when the number of PGS strands is less than half of the total deme size so $q=n/2$ in this work. Setting $q=n/2$ is an approximation in the DGS model since there can be more than two different gene sequences at the same time.

3. Discussion

Figure 2 shows CA, SGS, and DGS results for the value of I to be (a) 0.05 (strand gen)$^{-1}$, (b) 0.01 (strand gen)$^{-1}$, (c) 0.001 (strand gen)$^{-1}$, and (d) 0.0001 (strand gen)$^{-1}$. The SGS and DGS simulations begin with $q=10$, $\rho_q(0)=1$, and all other $\rho_k(0)=0$. The CA model required between 80 and 90 min of CPU time on a 2.6 GHz processor to complete the 10,000 gen calculation used in Figure 2(a) with FORTRAN. A week of CPU time was required to generate the results used in Figure 2(d). For comparison, the SGS and DGS results for Figure 2(a) and (d) were completed in 3 and 30 s, respectively, using SCILAB.

The CA model has already shown that the same genetic sequence can become the PGS multiple times when a deme's genetic sequence is polymorphic. Consequently, the rapid replacement of the PGS on the short time scale in the CA model represents a competition between genetic sequences and not the removal of a genetic sequence from the deme. The two ensemble models calculate the probability that a new PGS will remain the PGS for any specified duration. Both ensemble models predict the rapid replacement of an initial PGS as expected for a polymorphic deme where inversion size effect is the only cause of fitness differences between genetic sequences. A long exponential tail forms, following the polymorphic stage. This stage corresponds to a much more stable PGS when non-PGSs are rare. The polymorphic region is most visible in Figure 2(a), but all plots in Figure 2 show the result of genetic sequence competition as the fraction of demes in the ensemble with the original PGS decreases from 1 to approximately 0.3.

The rapid replacement of an initial PGS can also be explained by considering Equation (7). The first term in Equation (7) always acts to increase or decrease the PGS state, depending on whether the value of k is less than or more than j_{eq}, respectively. In the cases shown in Figure 2, it acts to move members of the ensemble from $\rho_q(0)=1$ to a distribution of non-zero $\rho_k(t)$ about $k=j_{eq}$. The second term describes the removal of all demes from the ensemble so it always acts to reduce the ensemble fraction with k PGS strands. In the first few generations, the second term in Equation (7) describes the rapid removal of demes until the first term has acted to transfer the mean density of PGS states from near q to near j_{eq}. The third term always acts to increase the ensemble fraction with k PGS strands, but has little effect on the ensemble prior to the first two terms having sufficient time to disperse the ensemble from the initial state.

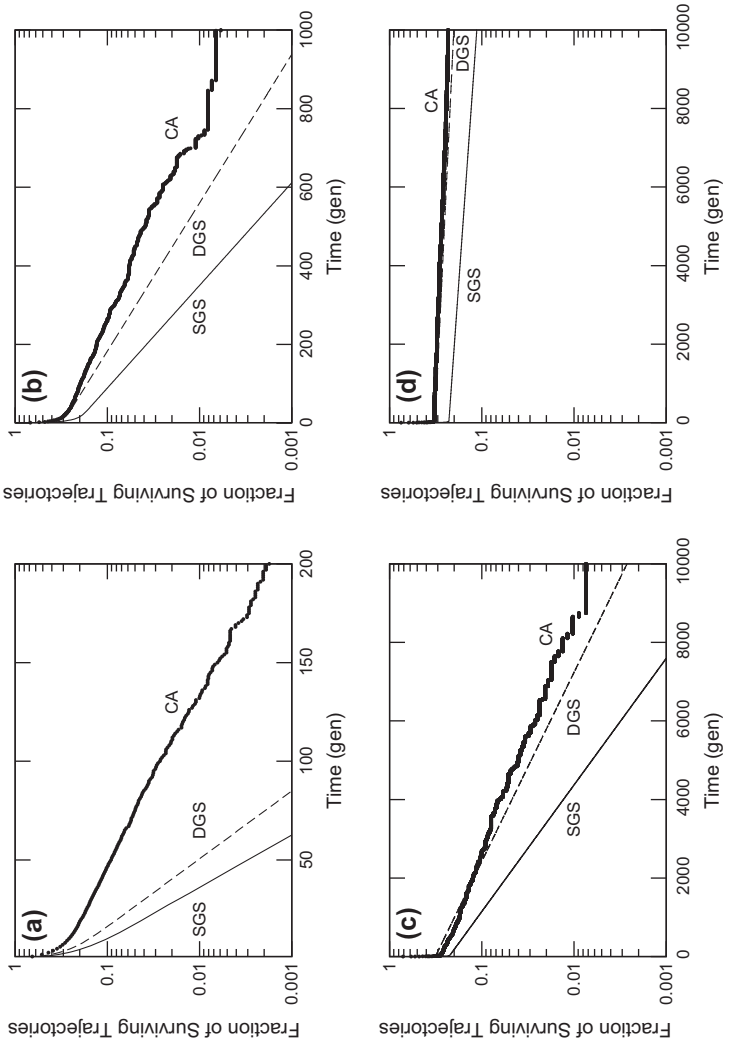

Figure 2. Fraction of surviving trajectories as a function of time for CA, SGS and DGS results for $I = (a)\ 0.05$ (strand gen)$^{-1}$, (b) 0.01 (strand gen)$^{-1}$, (c) 0.001 (strand gen)$^{-1}$, and (d) 0.0001 (strand gen)$^{-1}$. The thick broken curve is the CA data, the continuous line is the SGS model data, and the dashed line is the DGS model data.

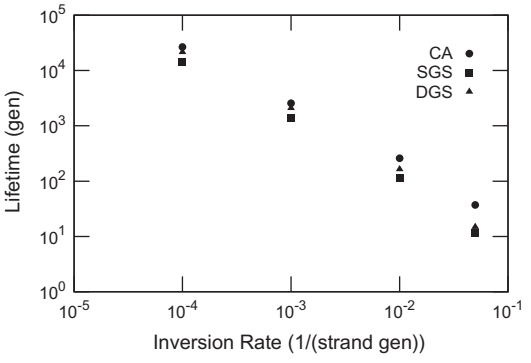

Figure 3. Comparison of lifetimes computed with the CA model and the SGS and DGS models using Equation (7) for $n = 20$.

The ensemble and CA results are not expected to agree exactly because the ensemble models consider the evolution from a single state until the PGS has been replaced, while the CA model follows all of the PGSs in the history of a single deme. Nevertheless, the ensemble and CA results should be in general agreement after the ensembles have evolved from the initial state, which usually takes less than 50 generations for the inversion rates used in this work. At this point, the ensembles have evolved away from demes that are all polymorphic to a density of states in which demes with rare inversions are common or even predominant as the inversion rate approaches zero. Figure 2 shows that the slopes at large times are similar between the ensemble models and the CA for each set of initial conditions even though the inversion rate changes by a factor of 500 from Figure 2(a)–(d).

Both ensemble models tend to underestimate the stability of the PGS for a deme size of 20 individuals. For example, the CA model converges to $\tau = 259$ gen, which is closer to $\tau = 164$ gen for the DGS model versus $\tau = 113$ gen for the SGS model when $I = 0.01$ (strand gen)$^{-1}$. Both SGS and DGS models only include stochastic fluctuations of the size of the PGS subpopulation relative to the total deme. This is valid for the SGS model, but an additional approximation in the DGS model. The DGS model assumes each non-PGS present in the deme has an occurrence frequency of one. This is accurate when inversions are rare but not when a deme is polymorphic.

The lifetime is well defined for exponential decay curves as in the tail regions of the curves in Figure 2. The lifetimes of a PGS for the CA, SGS and DGS model are provided in Figure 3 for the simulations shown in Figure 2. The lifetime of the PGS in the ensemble models becomes constant, when the relative distribution of states in the ensemble, $\rho_k(t)$, becomes constant, which requires that all $\Delta\rho_k(t)/\rho_k(t) = -1/\tau$ for t in the long tail region. The approximate lifetime can also be obtained analytically by solving a modified version of Equation (7). The second term in Equation (7), including the sign, can be conveniently written

$$\rho_k(t)\left(\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{(n-k)}-1\right), \tag{8}$$

and the first term within the outer parentheses in this expression can be combined with the third term in Equation (7) to yield

Table 2. Lifetimes for various inversion rates and deme sizes. Lifetimes for the CA, SGS and DGS models are included. (7) and (12) refer to the equation numbers in the text that are used for the lifetime calculation.

n	I((strand gen)$^{-1}$)	$\tau(10^3$ gen)				
		CA	$SGS(7)$	$DGS(7)$	$SGS(12)$	$DGS(12)$
20	0.05	0.0375	0.0115	0.0149	0.0213	0.0194
20	0.01	0.259	0.113	0.164	0.132	0.209
20	0.001	2.57	1.40	2.09	1.56	2.62
20	0.0001	26.5	14.3	21.4	15.9	26.7
40	0.001	6.98	4.00	11.3	3.92	10.1
80	0.001	19.6	44.4	526.	28.2	112

$$\sum_{i=0}^{n} \rho_i(t) \binom{n}{k} \left(\frac{i}{n}\right)^k \left(1 - \frac{i}{n}\right)^{(n-k)}. \tag{9}$$

Now it is useful to modify Equation (7) by multiplying the left side by $\rho_k(t)/\rho_k(t)$ and dividing through by $\rho_n(t)$ to obtain

$$-\frac{C_{kn}}{\tau} = \sum_{i=q}^{n} \sigma_{ki} C_{in} - C_{kn} + \sum_{i=q}^{n} B_{ki} C_{in}. \tag{10}$$

Here, $C_{kn} = \rho_k(t)/\rho_n(t)$, $C_{nn} = 1$, and the time dependent functionality for C_{kn} is not included since it approaches a constant value as time increases. Additionally, the ki terms in the last summation on the right side of Equation (7), excluding $\rho_i(t)$, are denoted by B_{ki} to obtain the simplified form of Equation (10). The value of τ in Equation (10) remains the negative, inverse slope of the straight line section of the curves in Figure 2, and can more generally be written as

$$\frac{1}{\tau} = \frac{\sum_{k=m}^{n} C_{kn} \sum_{i=0}^{q-1} \binom{q}{n} \left(\frac{k}{n}\right)^i \left(1 - \frac{k}{n}\right)^{(n-i)}}{\sum_{k=m}^{n} C_{kn}}, \tag{11}$$

where the inner sum is the cumulative distribution function for the probability of obtaining 0 to $q-1$ strands with the PGS starting with k PGS strands in a deme of size n.

Equation (10) is conveniently solved for the set of C_{kn} by moving the term on the left side to the right side and moving $B_{kn}C_{nn}$ from the summation in Equation (10) to the left side. Then the only non-zero term on the left side is $B_{nn}C_{nn}$. The resulting expression,

$$-B_{kn}C_{nn} = -\left(1 - \frac{1}{\tau}\right) C_{kn} + \sum_{i=q}^{n} \sigma_{ki} C_{in} + \sum_{i=q}^{n-1} B_{ki} C_{in}, \tag{12}$$

can be solved iteratively to obtain the values of C_{kn} and τ in approximately 0.7 s using SCILAB. Table 2 presents the PGS lifetime results for the same initial conditions as used for Figure 2 and is in qualitative agreement with the results from implementation of Equation (7) for both SGS and DGS models. Lifetimes for deme sizes of 40 and 80 individuals are included for an inversion rate of $I = 0.001$ (strand gen)$^{-1}$. The PGS lifetime calculation is most sensitive to the ensemble members in states with k near q, and it is insensitive to the value of C_{kn} for k near n. Figure 4(a) shows the $(n+1-q)$ values of C_{kn} for $I = 0.01$ (strand gen)$^{-1}$ and 0.001 (strand gen)$^{-1}$. A factor of ten decrease in inversion rate produces

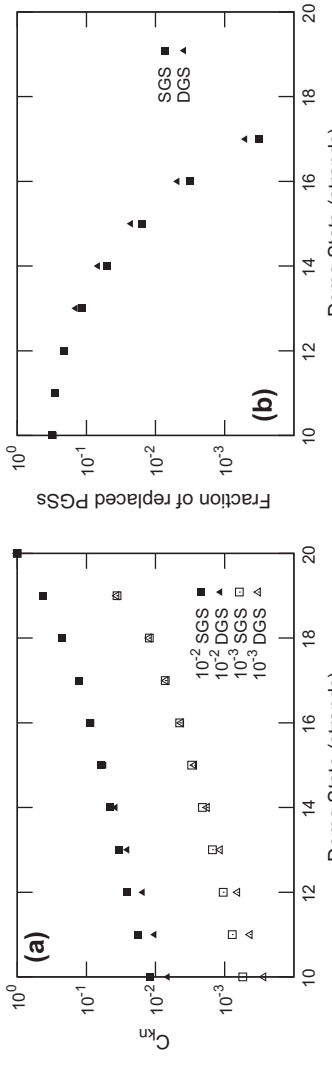

Figure 4. (a) C_{kn} and (b) fraction of replaced PGSs as a function of deme state. The key in (a) includes I in (strand gen)$^{-1}$.

a slightly greater decrease in the values of all C_{kn}, $k \neq n$, relative to C_{nn}. The relative decrease in values of C_{kn} remains nearly constant as k approaches q. Figure 4(b) shows the components of Equation (11) for $I = 0.01$ (strand gen)$^{-1}$. Demes whose states are defined by $q \leq k \leq q + 2$ account for 0.74 of PGS replacement in any generation. While the probability of PGS replacement for $k \geq n - 2$ is nearly zero.

Spirito (1992) calculated the probability of fixation of underdominant chromosomal inversions for a diploid deme, U, using several methods which were in qualitative agreement with each other. The results presented here for a haploid deme with 20 members can be compared to Spirito's results for a diploid deme with 10 members. Spirito reports rates for particular inversion sizes so we interpolated his results for the iteration of genotypic transition matrices to obtain a fixation probability of $U = 0.01017$ for $n = 10$ and $L = 0.14$ (pair gen)$^{-1}$. This was then converted to a lifetime according to

$$\tau = \frac{1}{nIU}, \tag{13}$$

which yields a calculated lifetime of 4.92×10^3 gen. Spirito's work includes the 0.5 probability that an inversion that becomes a PGS may not survive the polymorphic state, while our calculations are specifically for the rare state. To compare our results with Spirito's, Spirito's lifetime value can be reduced by a factor of two to account for the probability that the new PGS will not survive the polymorphic state, which gives a lifetime of 2.46×10^3 gen. This value compares well with a lifetime of 2.57×10^3 gen shown in Table 2 for the CA model. Similar calculations for $I = 0.01$ (strand gen)$^{-1}$ yield lifetimes of 10.8×10^3 gen and 162×10^3 gen for $n = 40$ and 80, respectively. Here the qualitative agreement between Spirito's results and lifetimes reported in Table 2 is not as good as n increases. Spirito's results, like our SGS and DGS model results, are based on approximations, so we expect the CA model results are more accurate.

Table 2 shows that the PGS stability has a non-linear dependence on the inversion rate that is significant for $I > 0.01$ (strand gen)$^{-1}$. The average fraction of the deme that will have a non-PGS genetic sequence at any time increases as the inversion rate increases. The CA results show that the PGS is replaced at a rate of once every 37 inversions at $I = 0.05$ (strand gen)$^{-1}$ compared to once every 51.4 inversions when $I = 0.001$ (strand gen)$^{-1}$. The non-linear effect is small enough in the CA model for $I \leq 0.01$ (strand gen)$^{-1}$ that it is unlikely to be observed in isolated demes. The same equations as used in this work can also be used to model gene sequence immigration into a deme. Immigration can occur at higher rates than inversion so non-linear effects should be easily observed in a controlled setting. For example, when there is one immigrant per generation in a deme size of 20.

4. Conclusion

The CA and ensemble models presented in this work provide useful ways to calculate the expected lifetime of a PGS when all genetic sequences are equally fit and mismatches due to inversion size and location are included. The models can be extended to incorporate other quantifiable fitness characteristics. The three models are in qualitative agreement with each other, although the ensemble models require far less computing resources. The CA model is useful because it can show an individual deme making the transition between polymorphic and rare states, while the SGS and DGS models show the ensemble behaviour. The CA model can be used to make predictions in either the polymorphic or rare states, while the ensemble models can be applied to rare states. These models can guide the design of experiments in controlled settings for small populations, although such experiments are likely to yield

results more quickly if conducted in the polymorphic state. The current ensemble models average the effect of the distribution of inversion sizes on the PGS lifetime. We plan to use the ensemble models to explore the predicted effect of inversion size on PGS lifetime and compare those results with the CA model. We also plan to generalize the models to include non-neutral inversions.

Acknowledgement

The author is thankful for the reviewers' comments. They have greatly improved the quality of this article.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Ayala, F. J., & Coluzzi, M. (2005). Chromosome speciation: Humans, Drosophila, and mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 6535–6542.

Beerli, P., & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics, 152*, 763–773.

Clark, B. K., Wabick, K. J., & Weidner, J. G. (2012). Inversion and crossover recombination contributions to the spacing between two functionally linked genes. *BioSystems, 109*, 169–178.

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

Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: Rearranging theory with pieces of evidence. *Trends in Ecology and Evolution, 25*, 660–669.

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh, 42*, 321–341.

Hedrick, P. W. (1981). The establishment of chromosomal variants. *Evolution, 35*, 322–332.

Hey, J., & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics, 167*, 747–760.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics, 47*, 713–719.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology, 8*, e1000501. doi:10.1371/journal.pbio.1000501

Lande, R. (1979). Effective deme sizes during long-term evolution estimated from rates of chromosomal rearrangement. *Evolution, 33*, 234–251.

Lee, J., Han, K., Meyer, T. J., Kim, H.-S., & Batzer, M. A. (2008). Chromosomal inversion between human and chimpanzee lineages caused by retrotransposons. *PLoS ONE, 3*, e4047. doi:10.1371/journal.pone.0004047

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology, 8*, e1000500. doi:10.1371/journal.pbio.1000500

Navarro, A., & Barton, N. H. (2003a). Chromosomal speciation and molecular divergence – Accelerated evolution in rearranged chromosomes. *Science, 300*, 321–324.

Navarro, A., & Barton, N. H. (2003b). Response to comment on 'Chromosomal speciation and molecular divergence – Accelerated evolution in rearranged chromosomes'. *Science, 302*, 988c.

Noor, M. A. F., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 12084–12088.

Ranz, J. M., Maurin, D., Chan, Y. S., von Grotthuss, M., Hillier, L. W., Roote, J., ... Bergman, C. M. (2007). Principles of genome evolution in the *Drosophila* melanogaster species group. *PLoS Biology, 5*, 1366–1381. doi:10.1371/journal.pbio.0050152

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology and Evolution, 16*, 351–358.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingston, K., Nakazato, ... Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science, 301*, 1211–1216.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications, 106*, 107–139.

Spirito, F. (1992). The exact values of the probability of fixation of underdominant chromosomal rearrangements. *Theoretical Population Biology, 41*, 111–120.

Spirito, F. (1998). The role of chromosomal change in speciation. In D. J. Howard & S. H. Berlocher (Eds.), *Endless forms: Species and speciation* (pp. 320–329). Oxford: Oxford University Press.

Wang, J., & Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics, 163*, 429–446.

White, M. J. D. (1978). *Modes of speciation*. San Francisco, CA: Freeman.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*, 97–159.

York, T. L., Durrett, R., & Nielsen, R. (2007). Dependence of paracentric inversion rate on tract length. *BMC Bioinformatics, 8*, 115–126. doi:10.1186/1471-2105-8-115

Appendix 1

In the limit of large inversion rate, inversions occur frequently enough that several other gene sequences exist in the deme along with the PGS. In the DGS model, we assume a different genetic sequence for all strands that do not have the PGS. This greatly simplifies the model but overestimates the stability of the PGS. The deme is divided into the j strands with the PGS and $n - j$ strands with other sequences, and all $n - j$ sequences are unique. The number of strands lost to inversion from the j strands with the PGS remains Ij. The recombination loss includes $Lj(1 - j/n)$ from each of the two subpopulations because of recombination between the subpopulations and $L(n - j)(1 - j/n - 1/n)$ from recombination within the subpopulation with $n - j$ unique genetic sequences. The total loss in population from recombination is

$$n_{\text{loss}} = L\left(1 - \frac{j}{n}\right)(n - j - 1)\,. \tag{A1}$$

Recombination losses to the reproducing population are replaced by allowing viable individuals to reproduce until the population is returned to n. The average change in the size of the population with the PGS is

$$\Gamma = -Ij - Lj\left(1 - \frac{j}{n}\right) + \left(L\left(1 - \frac{j}{n}\right)(n + j - 1)\right)\frac{\left(j - Ij - Lj\left(1 - \frac{j}{n}\right)\right)}{n - L\left(1 - \frac{j}{n}\right)(n + j - 1)}, \tag{A2}$$

where the first two terms represent the inversion and recombination reductions in the number of strands with the PGS, respectively. The third term accounts for the proportional replacement of strands lost from the entire deme. It is the product of the number of strands lost from the deme and the proportion of strands with non-PGSs after the inversion and recombination events described by the first two terms in the equation. Setting $\Gamma = 0$ yields a quadratic equation whose meaningful solution is

$$j_{eq} = \frac{n}{2}\left(\left(1 + \frac{1}{n}\right) + \sqrt{\left(1 - \frac{1}{n}\right)^2 - \frac{4I}{L}}\right). \tag{A3}$$

and provides the mean fraction of strands with the PGS when considering only the average effects of inversion, recombination, and deme size on reproduction. If $j_{eq} > n/2$, Equation (A3) shows that a

PGS will never be replaced if only average effects are considered, just as we observed for the SGS model discussed in the main text. If $I \ll L/4$, Equation (A3) can be approximated as

$$j_{eq} = n\left(1 - \frac{I}{L\left(1-\frac{1}{n}\right)}\right). \tag{A4}$$

Appendix 2

The average rate of unsuccessful recombination events depends on the average density of different inversion sizes in the deme with n total members and j members with the PGS. Let $\mu_i(t)$ be the density of all inversions of i genes with respect to the PGS. There are $(m+1-i)$ unique inversions of size i and the total number of possible inversions of $m(m+1)/2$, where the total number of genes is m and inversion loci are restricted to points between genes or at the ends of the strand of DNA. The rate of unsuccessful recombination events between the inverted strands and deme members with the PGS is $(i-1)/m$. This model assumes only one genetic sequence other than the PGS is present in the deme at any time. In a system with discrete generations, the change in $\mu_i(t)$ in one generation is

$$\Delta\mu_i(t) = 2I\left(\frac{j}{n}\right)\left(\frac{m+1-i}{m(m+1)}\right) - \left(\frac{j}{n}\right)\left(\frac{i-1}{m}\right)\mu_i(t), \tag{B1}$$

where I remains the inversion rate. The term with I is the rate at which new inversions of size i are produced while the term including $\mu_i(t)$ is the rate of loss of inversions of size i from the deme.

At steady state, $\Delta\mu_i(t) = 0$ and

$$\mu_i = 2I\left(\frac{m-(i-1)}{(m+1)(i-1)}\right), \tag{B2}$$

where the explicit time dependence has been dropped for the steady state value of μ_i. There are $(i-1)$ recombination locations for an inversion of size i, so the mean loss rate is

$$L = \frac{\langle i-1\rangle}{m}, \tag{B3}$$

where

$$\langle i-1\rangle = \frac{\sum_{i=2}^{m}\left(\frac{(m-(i-1))}{(m+1)(i-1)}(i-1)\right)}{\sum_{i=2}^{m}\left(\frac{(m-(i-1))}{(m+1)(i-1)}\right)}. \tag{B4}$$

The mean loss rate is independent of the inversion rate and the sum does not include $i = 1$ since an inversion of only one gene is meaningless in this model.