Gene conversion facilitates adaptive evolution on rugged fitness landscapes

Philip Bittihn*†,1 and Lev S. Tsimring*†,1

*BioCircuits Institute, UCSD, 9500 Gilman Dr., La Jolla, CA 92093-0328, †San Diego Center for Systems Biology, 9500 Gilman Dr., La Jolla, CA 92093-0688

ABSTRACT Gene conversion is a ubiquitous phenomenon that leads to the exchange of genetic information between homologous DNA regions and maintains co-evolving multi-gene families in most pro- and eukaryotic organisms. In this paper, we study its implications for the evolution of a single functional gene with a silenced duplicate, using two different models of evolution on rugged fitness landscapes. Our analytical and numerical results show that, by helping to circumvent valleys of low fitness, gene conversion with a passive duplicate gene can cause a significant speedup of adaptation which depends non-trivially on the frequency of gene conversion and the structure of the landscape. We find that stochastic effects due to finite population sizes further increase the likelihood of exploiting this evolutionary pathway. A universal feature appearing in both deterministic and stochastic analysis of our models is the existence of an optimal gene conversion rate which maximizes the speed of adaptation. Our results reveal the potential for duplicate genes to act as a “scratch paper” that frees evolution from being limited to strictly beneficial mutations in strongly selective environments.

KEYWORDS Gene conversion; Gene duplication; Pseudogenes; Fitness landscape; Adaptive evolution

copyright © 2017 by the Genetics Society of America
doi: 10.1534/genetics.117.300350
Manuscript compiled: Tuesday 3rd October, 2017%

It is widely believed that nonfunctionalization is by far the most common fate of of gene duplicates, with 90% of newly duplicate genes quickly turning into “pseudogenes” (Lynch and Conery 2000; Plata and Vitkup 2014). However, rendering a gene silent does not stop the evolution of its DNA sequence. Moreover, gene conversion, because it is linked to DNA replication, does not depend on the functionality of the gene product (Paulsson et al. 2017). Therefore, neutral mutations in a nonfunctionalized pseudogene can be transferred into the functional gene. If gene conversion introduces a lethal or deleterious mutation into a functional gene, this mutation would be quickly purged from the population; however, a beneficial mutation would have a high probability of fixation. Even a non-functional gene can therefore potentially be the source of mutations for adaptation.
are currently no models that address the advantage that gene conversion can provide to adaptive evolution in complex fitness landscapes.

Here, we study theoretically the role of gene conversion in adaptive evolution on non-trivial fitness landscapes for the simplest case of one pair of duplicate genes. In the first part of the paper, using a stochastic population dynamics model in a minimal non-trivial fitness landscape, we explore how gene conversion accelerates the fixation of a beneficial mutation that is only accessible via a deleterious intermediate mutation. In the second part of the paper, extending the concept to more complex genotypes and landscapes inspired by the classical NK-model for epistasis (Kauffman 1993), we then analyze how the rate of adaptation depends on the rates of gene conversion and mutation and the population size, enabling us to determine the conditions for the optimal gene conversion rate in rugged fitness landscapes.

Minimal model

Consider a pool of individuals with a genome consisting of a single “gene” that can only have three states, 0, 1, and 2. We assume a constant population size $N$, so the individuals reproduce with the replication rate equal to their corresponding fitness values and, to keep the population size constant, each replication event is accompanied by removal of a random individual from the population (Moran process). The population is initially monomorphic with all individuals in state 0. We assume that a single mutation can only switch the gene between states 0 and 1 or between states 1 and 2, but not between 0 to 2 directly. This would be the case, for example, if we consider a 2-bit string as a model gene, and a mutation can only flip one bit at a time (both one-bit mutations can be lumped in state 1). Evidently, only two sequential mutations may bring the system from the original state to the state with both bits flipped. We refer to the fitness values of states 0, 1, 2 as $F_0$, $F_1$, and $F_2$, respectively, and assume that they satisfy the condition $F_1 < F_0 < F_2$. Thus, to mutate from state 0 to a higher-fitness state 2, an individual has to cross a lower fitness “valley” of state 1. To describe gene conversion with a duplicate gene, we augment the “active” gene by a “passive” gene that can also be in any of the three states $\{0, 1, 2\}$, but the state of the passive copy does not affect the fitness of the individual. We denote the state of an organism in this $3 \times 3$ state space by $(i, j)$ where the two indices $i, j \in \{0, 1, 2\}$ correspond to the types of the active and passive genes, respectively (Figure 1b), meaning that the fitness of the state $(i, j)$ is $F_i$. We assume that mutations from state $(i, j)$ occur only during replication, and so they happen with propensity $\mu F_i$ (proportional to the fitness of the active copy) to each of the neighboring states, corresponding to forward and backward mutations in both active and passive copies. Gene conversion is modeled as copying the active gene into the passive gene $(i, j) \rightarrow (i, i)$ or back $(i, j) \rightarrow (j, j)$ with the rate $\alpha F_i$. Without loss of generality, we can rescale time to fix the wild-type fitness at 1 and take $F_0 = 1, F_1 = 1 - \delta, F_2 = 1 + s$. In the following, we will always assume $\mu, \alpha \ll s \ll 1$.

We performed stochastic simulations of this model with fixed selective advantage $s = 0.1$ and different values of $\delta, \mu, \alpha, N$ using the Gillespie algorithm (Gillespie 1977). We start simulations from all $N$ individuals localized in state $(0, 0)$ and observe the evolution of the distribution of the occupation numbers $X_{i,j}$ of each state $(i, j)$ over time. For mutation rate $\mu = 10^{-3}$ and population size $N = 800$ mutations occur relatively frequently. Without gene conversion ($\alpha = 0$) the state of the passive gene is

---

**Figure 1** (a) Cartoon of the gene duplication and conversion. (b) Schematic diagram of the minimal $3 \times 3$ model. Short bidirectional arrows denote mutations and green arrows denote gene conversions.
irrelevant, and evolution proceeds via regular mutations along the state of the active gene (Figure 2a). The population spreads from state 0 to state 1 and then to 2 where it eventually localizes. The efficiency of this process depends critically on the fitness of the intermediate state 1, and for large $\delta$ (small $F_1$) is extremely low. However, for finite gene conversion rate ($\alpha > 0$), the population can reach state 2 for the active gene even across a deep fitness valley ($\delta = 1$), as shown in Figure 2b. The population spreads transiently to state (0,2) (light blue) by mutations $(0,0) \rightarrow (0,1) \rightarrow (0,2)$ and then crosses the fitness valley directly to state (2,2) (bright green) by gene conversion.

If the flux of mutations is sufficiently small, the dynamics look markedly different (Figure 2c and d), as a polymorphic population similar to Figure 2a and b cannot be maintained. Instead, each mutation either gets fixed or is lost, i.e. the population homogenizes before the next mutation occurs, leading to “switching” dynamics. If a weakly deleterious mutation into state 1 gets fixed by random genetic drift (red color), the population can reach state 2 even when $\alpha = 0$ (Figure 2c). For $\alpha > 0$ and $\delta = 1$ (Figure 2d), the population reaches state (2,2) (bright green) via conversion after a neutral mutation to state (0,2) becomes fixed in the population (light blue color).

As shown by Weissman et al. (2009), the way a population crosses intermediate states to a final beneficial mutant (in our case either along the direct or along the gene conversion path) can differ significantly depending on population size $N$ and the depth of the fitness valley $\delta$. It can range from (1) a sequence of fixation events at very small population sizes, with the population occupying essentially a single state at a time (Figure 2c,d), to (2) “stochastic tunneling” at intermediate population sizes, when intermediate states can be occupied long enough to result in the production of a beneficial mutant (Figure 2a,b), to (3) completely deterministic continuous generation of mutants at very large population sizes. A number of additional mixed cases can be expected when either the two paths (direct or via gene conversion) are in different regimes or individual transitions within the paths have to be treated with different approximations. While all of these mixed regimes can in principle be analyzed, we will concentrate on the three pure regimes mentioned above. The exact population size requirements for each case will be specified below.

**Figure 2** Stochastic simulations of the minimal gene conversion model. Stacked areas of different colors represent the number of individuals in each state according to the lookup table in panel (a). Colors blue, red and green correspond to the state of the active gene determining the fitness, different shades of the same color indicate the state of the passive gene. Parameters: (a) $N = 800$, $\delta = 0.03$, $\mu = 10^{-3}$, $\alpha = 0$. (b) $N = 800$, $\delta = 1$, $\mu = 10^{-3}$, $\alpha = 10^{-3}$. (c) $N = 100$, $\delta = 0.01$, $\mu = 10^{-4}$, $\alpha = 0$. (d) $N = 100$, $\delta = 1$, $\mu = 10^{-4}$, $\alpha = 10^{-4}$. In all panels, $s = 0.1$.

We can apply this approach to our problem by computing the fixation times of the beneficial state 2 for evolution via two separate paths which in the first approximation (i.e. for small mutation and conversion rates) can be considered independent. The direct path (0 $\rightarrow$ 1 $\rightarrow$ 2) is essentially the valley crossing problem of Weissman et al. (2009) for $K = 2$ (with one deleterious intermediate mutation), with a few minor differences. First, our model postulates that the fitness of an individual characterizes its total division rate that includes both normal symmetric binary divisions and mutations. Accordingly, the mutation rate of the individual is equal to the product of the nominal per-generation mutation rate $\mu$ times its fitness. Additionally, we take into account back mutations 1 $\rightarrow$ 0 and 2 $\rightarrow$ 1 with the same rates as the forward mutations. Still, the first-step analysis can be applied here, and in the limit of strongly deleterious intermediate state ($\delta \gg \sqrt{(1-\delta)\mu s}$), the rate of producing single mutants in state 1 is $N\mu$ and the probability that this mutant is successful is $p_1 = (1-\delta)\mu s/\delta$, so the approximate expression for the fixation time $T_1 = 1/(\mu N p_1)$ of the direct path reads

$$T_1 \approx \frac{\delta}{N\mu^2 s (1-\delta)}$$

in agreement with previous studies of fitness valley crossing by Innan and Stephan (2001); Weissman et al. (2009) (see File S1 for more details). As expected, the fixation time is inversely proportional to the population size and strongly dependent on the mutation rate $\mu$ and the selective advantage of the beneficial state $s$. 

**Stochastic tunneling**

In this regime, the mutant subpopulation remains much smaller than $N$ until the final stage when beneficial mutants are established in a population and finally grow to fixation (see Figure 2a,b). The adaptive evolution proceeds via stochastic tunneling through the low-occupancy intermediate states. Weissman et al. (2009) presented the most comprehensive theoretical description of stochastic tunneling to a beneficial state with $K$ mutations through a low-fitness valley of $K-1$ deleterious or neutral mutations. Their approach is based on the notion of a “successful” individual. An individual is considered successful if one of its descendants will mutate into the final (beneficial) state and get fixed in the population. We define the fixation time of the beneficial mutation as the time when half of the population has reached the state 2. In the stochastic tunneling regime, the dominant contribution to this fixation time is the time until a successful individual is produced. This time is inversely proportional to the rate of producing a single mutant and to the probability $p_1$ that this mutant is successful. The latter can be computed by the first-step analysis described in Appendix C of Weissman et al. (2009).

Gene conversion facilitates adaptive evolution on rugged fitness landscapes
The second path that involves gene conversion can be exactly isolated by assuming $F_1 = 0$ ($\delta \to 1$). In this case the mutants in the intermediate state 1 do not reproduce and are effectively excluded from the evolutionary process. Thus, the only path to the beneficial state 2 is $(0,0) \to (0,1) \to (0,2) \to (2,2)$, where the first two steps are neutral mutations in the passive copy, and the last step is the gene conversion from the passive to the active copy. Although this regime is similar to the case $K = 3$ with two neutral intermediate mutations in Weissman et al. (2009), the inclusion of back mutations $((0,1) \to (0,0)$ and $(0,2) \to (0,1)$) and—more importantly—back conversions $((0,1) \to (0,0), (0,1) \to (1,1), (0,2) \to (0,0))$ leads to significantly different results: Using the first-step method again, we can obtain the full solution for the probability $p_{01}$ that the $(0,1)$ mutant is successful (see File S1) and derive the fixation time $T_2$ as $(N\mu p_{01})^{-1}$. This fixation time is plotted in Figure 3a as a function of $\alpha$ for fixed $N$ and $\mu$. As seen from this figure, it is non-monotonic, i.e. decreases at small $\alpha$ and increases at large $\alpha$, and there is an optimal $\alpha_m \approx 2^{-4/3}(s\mu^2)^{1/3}$ for which the fixation time $T_2$ is minimized. It is easy to understand the origin of this optimum: for small $\alpha$ the gene conversion proceeds very slowly, and for large $\alpha$ the process is hampered by the strong flux of mutants from states $(0,1)$ and $(0,2)$ back to state $(0,0)$ and to the low-fit state $(1,1)$, thus reducing the influx into state $(2,2)$. In other words: Gene conversion enables the transfer of the beneficial genotype from the passive to the active gene and is thus beneficial, but only up to a certain rate. Further increasing the gene conversion rate prevents the passive copy from reaching the beneficial state, because it is either replaced by the active copy or a deleterious intermediate state of the passive copy is transferred to the active gene. This trade-off, which is due to the bi-directionality of gene conversion, is a universal effect encountered in all the regimes and scenarios considered in this study.

The approximate formulas for $p_{01}$ for small and large $\alpha$ can be obtained in closed form (see File S1),

$$p_{01} \approx \begin{cases} \sqrt{\mu_s \alpha}, & \alpha \ll (s\mu^2)^{1/3} \\ \frac{\mu}{2 \sqrt{\pi}}, & \alpha \gg (s\mu^2)^{1/3} \end{cases}$$

which yields the following expressions for the fixation time $T_2$ of the gene conversion path,

$$T_2 \approx \begin{cases} 1/[N\mu^{3/2}(as)^{1/4}], & \alpha \ll (s\mu^2)^{1/3} \\ 2\alpha^{1/2}/[N\mu^2s^{1/2}], & \alpha > (s\mu^2)^{1/3} \end{cases}$$

Note that the smaller-$\alpha$ regime corresponds to the approximation derived by Weissman et al. (2009) whereas the second, larger-$\alpha$ regime is a result of back conversions and transitions to the deleterious state $(1,1)$ that lead to a reduction of the overall success probability of mutants in state $(0,1)$.

For small but non-zero fitness $F_1$ of the intermediate mutant, when both direct and gene conversion paths can lead to fixation of state 2 in comparable times, the overall fixation time can be found as $T = (T_1^{-1} + T_2^{-1})^{-1}$, since these paths are non-overlapping for large $\delta$ and small $\alpha$. An example for the overall fixation time of the beneficial mutation for fixed $N$ and $\mu$ is

1 The two cases for large and small $\alpha$ connect at $\alpha_m = 2^{-4/3}(s\mu^2)^{1/3}$. For specifying approximate regime boundaries, we omit the numerical factor $2^{-4/3}$.

2 Note also that there is a third regime for very small $\alpha \ll \mu^2/s$ (see File S1) which is usually irrelevant because $p_{01}$ drops very quickly in this regime and the direct path will consequently provide a shorter fixation time.
shown in Figure 3b for a range of conversion rates. For large $a$, the approximation $T = (T_1^{-1} + T_2^{-1})^{-1}$ slightly overestimates the fixation time because of the neglected interaction between the two paths: Conversions from (0,1) to the deleterious state (1,1) can facilitate tunneling along the direct path to the better fitness state (2,1) instead of simply representing an effective reduction in fitness.

Substituting the optimal $a_m$ into Equation (3) shows that the minimal fixation time achievable via gene conversion is $T_{2,\text{min}} \approx 2^{1/3} (N\delta^{1/3} \mu^{2/3})^{-1}$. Comparing this expression with the fixation time $T_1$ via the direct path, Equation (1), we observe that it is also inversely proportional to $N$, but depends more weakly on the mutation rate $\mu$ and much more weakly on the selective advantage $s$. Therefore, as shown in Figure 3c and d, the speedup through gene conversion in the stochastic tunneling regime is expected to be larger for smaller $\mu$ and $s$, as $T_{2,\text{min}}$ grows more slowly than $T_1$, while it does not change with $N$. The scaling behavior is reversed and becomes more favorable for the direct path if the fitness valley is effectively neutral ($\delta \ll \sqrt{(1-\delta)\mu s}$, see File S1), which is intuitive since the conversion path involves an extra step.

The applicability of the stochastic tunneling calculation requires that mutant sub-populations remain smaller than $N$ at all times except for the final fixation state of the beneficial mutant.

For the direct path, the sufficient condition for this is $p_1 \gg p_1$, where $p_1$ is the probability that a successful single mutant, and $p_1$ is the probability of fixation in state 1. For the strongly deleterious intermediate state ($\delta \gg 2\sqrt{(1-\delta)\mu s}$), we use Kimura’s diffusion approximation to the fixation probability (Kimura 1962; Otto and Whitlock 2006) for a mutation with selective advantage $s$

$$p(s) = \frac{1 - \exp(-s)}{1 - \exp(-Ns)}.$$  \tag{4}

It is exponentially small for negative $s$ (deleterious mutations), equal to $1/N$ for $s=0$, and is $\approx s$ for small $s > 0$ and large $Ns$. In this case, therefore, $p_1 = p(-\delta)$. This translates to the condition

$$N \gg \delta^{-1} \log \left(1 + \delta \frac{e^{-\delta} - 1}{\mu(1-\delta)s}\right). \tag{5}$$

which corresponds to Equation (31) of Weissman et al. (2009).

For the gene conversion path, the corresponding condition is $p_{01} \gg p_{01}$, where $p_{01}$ is the fixation probability in state (0,1).

For this state, gene conversion to the states (1,1) and (0,0) effectively reduces the division rate by a factor $1 - 2\alpha$, so its fixation probability is given by $p_{01} = p(-\delta)$. Substituting $p_{01}$ from Equation (2), we arrive at

$$N \gg \begin{cases} 
\frac{1}{\sqrt{\mu \sqrt{s}}}, & a \ll (s \mu)^{1/3} \\
\frac{1}{s} \log \left(1 + \frac{s}{\sqrt{2}} \sqrt{\frac{a}{s}}\right), & a \gg (s \mu)^{1/3},
\end{cases} \tag{6}$$

(see File S1). In Figure 3b, the shaded area indicates the values of $a$ at which this condition is violated for the gene conversion path, in this case only $N < 1/\sqrt{\mu \sqrt{s}}$. However, the direct path dominates the dynamics in this area and so, despite this, the quantitative deviations of the theoretical predictions remain invisible.

Besides a minimum population size, another requirement for the validity of the stochastic tunneling approximation is that $N$ is not too large, so the lineages of successful first mutants do not overlap, and the first one always leads to fixation. The sufficient condition for this to hold is

$$N \ll \delta^2 / [s \mu^2 (1-\delta)] \tag{7}$$

for the direct path with strongly deleterious intermediate state 1 and

$$N \ll \begin{cases} 
\frac{1}{s}, & a \ll (s \mu)^{1/3}, \\
\frac{4a s^{1/2}}{p \mu^{3/2}}, & a \gg (s \mu)^{1/3},
\end{cases} \tag{8}$$

for the gene conversion path (see File S1 for details).

**Continuous evolution in large populations**

When the conditions discussed at the end of the previous section are violated, the stochastic tunneling approximation is no longer valid. For very large populations (the validity conditions will be specified below), the population continuously spreads to neighboring states, and we can analytically describe the evolutionary dynamics by using deterministic mass-action equations for the occupation numbers $X_{ij}$:

$$\dot{X}_{ij} = F_i X_{ij} - d X_{ij} + \mu F_{i-1} X_{i-1,j} + \mu F_i X_{i-1,j} + \mu F_{i+1} X_{i+1,j} + \mu F_i X_{i,j+1} - (4 - \delta s) \mu F_i X_{i,j} - 2 \alpha \mu F_i X_{i,j} + \alpha \sum_{k=0}^{2} (F_i X_{ik} + F_k X_{ik}) \delta_{ij}, \{i,j\} = 0,1,2 \tag{9}$$

where $\delta_{ij}$ is the Kronecker delta, and $X_{ij} = 0$ for either $i$ or $j < 0$ or $2$. To mimic the Moran process deterministically, we set the death rate $d$ equal to the current average fitness of the whole population, $d = \sum_{j=0}^{2} F_i X_{ij} / N$, where $N = \sum_{i,j=0}^{2} X_{ij}$ is the constant population size.

For the model (9) the fixation time can be found analytically in two limiting cases using Laplace transforms (see File S1 for derivations). In the absence of gene conversion ($\alpha = 0$) and for small mutation rate $\mu \ll s$ the fixation time via the direct path is

$$T_1 \approx \frac{1}{s} \log \left[ \frac{s(\delta + s)}{2 \mu^2 (1-\delta)} \right] \tag{10}$$

In the other limiting case we again consider only the evolutionary path to state 2 via gene conversion by assuming $\alpha > 0$ but taking $F_1 = 0$. In this case the fixation time is given by the formula

$$T_2 \approx \frac{1}{s} \log \left[ \frac{s(2a + s)^2}{2 \mu^2} \right] \tag{11}$$

According to Equation (11), the switching time $T_2$ is again a non-monotonous function of $\alpha$ with a minimum at $a_m = s/2$. We can also find the value of $\alpha$ at which $T_1 = T_2$, as a root of the quadratic equation $(1-\delta)(s + 2\alpha)^2 = \alpha(s + \delta)$. For small $\alpha \ll s$ the solution is

$$\alpha_c = \frac{(1-\delta)s^2}{s + \delta}. \tag{12}$$

For $\alpha \ll \alpha_c$, the effect of gene conversion is negligible, and the fixation time is given by $T_1$ that depends on $F_1$. For $\alpha \gg \alpha_c$, gene conversion dominates the evolution, and the fixation time is given by $T_2$ that is independent of $\delta$. Interestingly, the crossover point $\alpha_c$ is independent of the mutation rate $\mu$. 

Gene conversion facilitates adaptive evolution on rugged fitness landscapes
Simulations of the full deterministic model (9) are in agreement with these results. Figure 4 shows the fixation time of state 2 as a function of $\alpha$ for different values of $\delta$ and $\mu$. As predicted by the theory, the fixation time is approximately $T_2$ for large enough $\alpha$, but at a certain $\alpha_c$ it crosses over to the $\alpha$-independent, but $\delta$-dependent value $T_1$. The fixation times deviate somewhat from the analytical estimates since we assume small $X_{ij}$ for all $(i,j) \neq (0,0)$ to derive them (see File S1 for details), but use a numerical fixation criterion of 50% for state 2. However, the relative numerical error of our analytical predictions is small and the parameter dependence of the fixation times is accurately described by the theory.

Intuitively, a smaller mutation rate $\mu$ leads to longer fixation times across all $\alpha$ (Figure 4b), as can also be seen directly from Eqs. (10) and (11). For fixed $s$, the crossover value $\alpha_c$ decreases monotonically with increasing $\delta$, and so the range of conversion rates $\alpha$ with shortened fixation times (when $T_2 < T_1$) increases (cf. Figure 4a and Figure 4b). For $\alpha < \alpha_c$, the high fitness state is reached directly through mutations, such that state (2,0) is populated first (Figure 4c). In contrast, for $\alpha > \alpha_c$, a state with beneficial fitness is reached by gene conversion into state (2,2) (Figure 4d), leading to reduced fixation time (Figure 4a).

The deterministic description presented in this section is applicable if for every transition between states included in the model $p_{a\rightarrow b}X_{ab}(t_e) \gg 1$, where $p_{a\rightarrow b}$ is the probability of transition from state $a$ to state $b$ and $X_{ab}(t_e)$ is number of individuals of type $a$ at the time $t_e$ at which the population of the beneficial mutants is established, i.e. reaches $O(s^{-1})$, see Fisher (2007); Weissman et al. (2009). For the direct path with strongly deleterious state 1, the most stringent limitation is imposed by the transition from state 1 to 2, for which $p_{1\rightarrow 2} = \mu(1-\delta)$ and $X_1(t_e) = \mu N/\delta$, so the deterministic description is valid for $N \gg \delta/\mu^2(1-\delta)$ (cf. Equation (36) of Weissman et al. (2009)). For the gene conversion path, the most stringent limitation is imposed by the transition from state (0,2) to (2,2), for which $p_{02\rightarrow 22} = \alpha$ and $X_{02}(t_e) = N\mu^2s^{-2} \log^2(s^2/N\alpha^2\mu^2)$ (see File S1).

Then, the condition of applicability of the fully deterministic description of the gene conversion path is

$$N \gg \frac{s^2}{\alpha \mu^2}.$$  \hspace{1cm} (13)

It is easy to see that there is a gap between this fully deterministic regime and the stochastic tunneling regime ($N \ll \mu^{-1}$). In this gap, various intermediate regimes apply in which some transitions can be treated deterministically and while others still exhibit stochastic tunneling. The corresponding theoretical approximations can be developed for these scenarios, but since there are many cases to consider for different combinations of parameters, we will not describe them here.

**Monoclonal regime in small populations**

If $N$ becomes so small that the conditions $p_1 \gg \rho_1$ (direct path) or $p_{01} \gg \rho_{01}$ (conversion path) for the applicability of the stochastic tunneling calculation are violated, the intermediate mutants may get fixed before giving rise to the beneficial mutant. For the direct path, the population size has to be extremely small for the strongly deleterious mutation to have a chance of getting fixed before a successful single mutant is produced by tunneling, because $\rho_1$ is exponentially small for $\delta N \gg 1$. However, if...
Nonetheless \( p_1 \ll \rho_1 \), i.e.

\[
N \ll \delta^{-1} \log \left( 1 + \delta \frac{\mu^3 - 1}{\mu (1 - \delta)s} \right) \tag{14}
\]

(The reverse of Equation (5)) is fulfilled, the intermediate state \((1,^*)\) will always get fixed before the mutation to state \((2,^*)\) happens. We call this the *monoclonal regime*.

For the conversion path with two intermediate mutations, the condition for fully monoclonal regime is not simply the reverse of the stochastic tunneling condition \( p_{01} \ll \rho_{01} \) since it only gives the estimation of state \((0,1)\) but not of state \((0,2)\). The condition for the latter is \( p_{02} \ll \rho_{02} \) which is always more stringent. Substituting \( p_{02} = \sqrt{\delta s} \) and \( \rho_{02} = p(-\alpha) \), we obtain

\[
N \ll \frac{1}{\sqrt{\delta s}} \tag{15}
\]

(see File S1 for details). In the gap between this condition and (6), there is a semi-monoclonal regime, where \((0,1)\) mutants get fixed and the population then stochastically tunnels through \((0,2)\) to the final beneficial state \((2,2)\). As mentioned in the introduction, we ignore these mixed scenarios in this study.

If we assume that the population gets fixed at every intermediate state along the way before proceeding (or possibly reverting due to back mutations or conversions), we can convert the full minimal model to a continuous-time Markov chain. In this *monoclonal approximation*\(^3\), different states represent homogeneous populations with a certain genotype and the transition probabilities between states depend on the corresponding mutation rates and fixation probabilities.

The fixation probability of a mutation from a state with fitness \( F_i \) to a state with fitness \( F_j \) is \( p(s_{ij}) \) from Equation (4) (diffusion approximation\(^4\)), where \( s_{ij} = (F_i - F_j) / F_j \) is the fitness advantage of state \( j \) over state \( i \). Multiplying this fixation probability by the corresponding mutation or conversion rates, we obtain transition rates between the different monoclonal states of the system. For the gene conversion path, the neutral mutations between states \((0,0)\), \((0,1)\) and \((0,2)\) occur with rate \( \mu N \) and become fixed with probability \( 1/N \), leading to the transition rate \( \mu \). Similarly, neutral conversions from \((0,1)\) and \((0,2)\) back to \((0,0)\) have transition probability \( \alpha \). The transition rate for the conversion from state \((0,2)\) to state \((2,2)\) is \( \alpha' = \alpha N p(s) \). The subsequent transitions among states \((2,2)\), \((2,1)\), and \((2,0)\) are not important since we are only interested in fixation of the active gene, thus we can lump these three states into one state \((2,^*)\).

Thus, the evolution along the gene conversion path can be reduced to a 4-state Markov chain with the single absorbing state \((2,^*)\), see Figure 5. The average time to reach state \((2,2)\) from state \((0,0)\) can be computed as

\[
T_2 = \frac{a^2 + a' a + 4 \mu a + 3 \mu^2 + 3a' \mu}{a' \mu^2} \tag{16}
\]

(see File S1). Numerical simulations of the full stochastic model agree with Equation (16) for small enough population sizes (Figure 6a). The time to fixation in a high-fitness state \((2,2)\) is

\[
T_2 = \frac{a^2 + a' a + 4 \mu a + 3 \mu^2 + 3a' \mu}{a' \mu^2} \tag{16}
\]

\[
(0,0) \xrightarrow{\mu} (0,1) \xrightarrow{\mu} (0,2) \xrightarrow{\alpha'} (2,^*)
\]

*Figure 5* Markov chain for an evolving population in the absence of the direct path (\( \delta = 1 \)) in the monoclonal approximation. Black arrows denote ordinary mutations and green arrows denote gene conversions.

Again, if the direct path and the conversion path are assumed to be independent, the two approximations can be combined to yield an overall fixation time of \( T = (T_1^{-1} + T_2^{-1})^{-1} \). A comparison with numerical simulations for different depths of the fitness valley is shown in Figure 6b. For these parameters, both the gene conversion path and the direct path are in the monoclonal regime. For large enough \( \delta \), the fixation time stays close to that predicted for the gene conversion path alone, as the direct path is strongly suppressed for deep fitness valleys.

It is interesting to note that substituting the optimal \( \alpha_m \) into Equation (16) yields a minimum achievable \( T_2 \) for the gene conversion path that is proportional to \( 1/\mu \). Since \( T_1 \) follows the same scaling law, the speed-up factor provided by gene conversion with an optimal rate is independent of \( \mu \) as long as both paths fulfill the conditions for the applicability of the monoclonal approximation, which is in contrast to the behavior of the stochastic tunneling approximations.

**Behavior across different regimes**

Figure 6c shows the behavior of the fixation time along the gene conversion path for increasing population size \( N \) and otherwise fixed parameters. For very small \( N \), the monoclonal approximation for the fixation time applies (black line). The stochastic tunneling approximation (dashed-dotted lines) correctly predicts the initial behavior when the population size leaves the monoclonal regime. For larger population sizes, the numerical results then start deviating from this prediction before eventually converging to the deterministic prediction (dashed lines).

The fixation time in the deterministic theory is independent of the population size, as long as it is large enough. However, when the population size drops below the applicability of the deterministic predictions, the actual fixation time may strongly deviate from these predictions. Figure 6d shows a comparison of the deterministic theory to deterministic and stochastic simulations for a parameter combination that is on the border of the deterministic regime. For relatively shallow fitness valley \( \delta = 0.5 \) (red), both deterministic and stochastic simulations are in agreement with deterministic theory: for small \( \alpha \) (where \( T_1 < T_2 \)), the data follows \( T_1 \) estimate, and for large \( \alpha \) (where

\[\text{Gene conversion facilitates adaptive evolution on rugged fitness landscapes}\]
Figure 6 Simulations of the full stochastic model (Figure 1b) with \( s = 0.1 \). Fixation times \( T \) obtained as averages from 5000 independent simulations unless stated otherwise. Symbols represent numerical simulations. (a) Fixation times for \( \mu = 10^{-7} \) and \( \delta = 1 \) as a function of \( \alpha \) for different \( N \). Lines represent the monoclonal approximation, (16). (b) Fixation times for \( \mu = 10^{-5} \) and \( N = 50 \) compared to the monoclonal approximation for the gene conversion path (16) (black line), for different values of \( \delta \). The colored lines are analytical predictions combining the monoclonal approximation for the direct and the conversion path according to \( T = (T^{-1} + T^{-1})^{-1} \). (c) Fixation times for \( \delta = 1 \) as a function of \( N \) for two different combinations of \( \alpha \) and \( \mu \), compared to the monoclonal approximation (16) (solid line), the tunneling approximation (dashed-dotted lines) and the deterministic approximation (11) (horizontal dashed lines). Averages for larger \( N \) were obtained from progressively fewer numerical simulations. (d) Fixation times for \( \mu = 10^{-3} \), large population size \( N = 10^6 \) and two different \( \delta \) from 50 simulations (circles) compared to deterministic simulations (squares) and formulas (10) (color lines) and (11) (black line).

As in our minimal model, we assume that the “gene” has undergone duplication and consider an extended genotype, augmenting the active \( N \)-bits gene by an additional “passive region” of equal length \( N \) which does not impact fitness. Mutations correspond to the flipping of a single bit, while gene conversion is implemented by replacing the entire passive region with the sequence in the active region or vice versa. The dynamics of evolu-
Gene conversion facilitates adaptive evolution on rugged fitness landscapes

Figure 7 Summary of different approximations. (a) Regimes of validity for the different approximations of the gene conversion path in the minimal model in terms of the population size $N$ and the conversion rate $\alpha$ for fixed $\mu = 10^{-6}$ and $s = 0.1$. The vertical dashed line indicates $\alpha = 2^{-4/3}(s\mu^2)^{1/3}$. (b) Speed-up $T_2/T_1$ provided by gene conversion across different population sizes for fixed $\mu = 10^{-6}$, $s = 0.1$ and $\delta = 0.1$ (the individual fixation times of the two paths can be found in Fig. S3). Results for each regime are plotted where the population size requirements for both the direct and the gene conversion path are fulfilled and both paths are in the same regime, except for the dashed red line labeled “mixed”, where the direct path is in the tunneling regime and the gene conversion path is in the monoclonal regime (cf. Fig. S3). Calculations are shown for a fixed gene conversion rate $\alpha_0 = (16 \cdot 10^{13})^{-1/3}$ (the optimum in the tunneling regime; blue lines) and the optimal $\alpha$ for each regime and population size (red lines).

Evolution and adaptation is then determined by the characteristic mutation rate per bit per individual $\mu$ and the gene conversion rate $\alpha$ per individual. We assume here that the population size $N$ is small enough such that populations are monoclonal. Furthermore, we assume strong selection, such that beneficial mutations and conversions always lead to fixation, while deleterious mutations and conversions are always rejected. For neutral mutations like those in the passive region, the fixation probability is again $1/N$. These two simplifications—monoclonal population and strong selection—allow us to efficiently obtain statistics over many possible paths in several randomly generated fitness landscapes and to observe the impact of gene conversion, since evolution away from a genotype without gene conversion is prohibited once it has reached a fitness peak.

We used the Gillespie algorithm (Gillespie 1977) to simulate evolution of a monoclonal population in this landscape after placing it in a random fitness peak and trace its fitness as it evolves over time. Figure 8a shows a sample trajectory of the actual fitness for a population initially trapped in a local fitness peak (thick solid black line) and the corresponding potential fitness of passive region (thin solid green line). The symbols mark mutations and gene conversion events that change the potential fitness of the passive region. This example includes a single fitness improvement step at $t \approx 5.15$ due to gene conversion from the passive region to the active one, after a beneficial mutation is discovered as a result of two consecutive mutations in the passive region. Note that immediately after this conversion, the actual fitness instantly climbs slightly above the potential fitness because the active region continues to evolve to the nearest fitness peak. While the actual fitness of individual trajectories is discontinuous, the ensemble-averaged fitness increases continuously and nearly linearly at small $t$ (Figure 8b,c). Therefore, we can use its initial slope $r$ (the adaptation rate) as a quantitative measure for the advantage that gene conversion confers on the adaptability of the population.

Figure 8 Fitness evolution in a rugged landscape. (a) Example trajectory of fitness of a single monoclonal population in landscape with $N = 20$, $X = 19$, $N_0 = 10$ and $N = 10^5$ and $\alpha/\mu = 10$. Thick black line indicates actual fitness of the population, thin green line shows the potential fitness of the passive region. (b) Ensemble-averaged fitness over time for select conversion rates $\alpha$ (solid lines) and control simulations without gene conversion (dashed lines). Other parameters are identical to those in (a). Fitness values were averaged over 1000 populations and 20 landscapes. (c) Initial average fitness trajectories for a range of different $\alpha$. Colored sections indicate the intervals which were used to detect the initial rate of fitness increase $r$. The color corresponds to the value of $\alpha$, cf. (d). (d) Initial rate of fitness increase $r$ extracted from the data in (c). Dashed lines indicate the approximations of Eqs. (18) (blue), (19) (yellow) and (20) (red).
**Trade-off between mutations and conversions leads to optimal adaptation**

Figure 8d shows that, for the minimal model, the rate of fitness improvement also has a maximum at some intermediate value of $\alpha$. For large $\alpha$, the passive copy rarely evolves more than 1 bit-flip away from the genotype in the active region, because it is reset very frequently, and so the adaptation rate $r$ decreases towards larger $\alpha$ (regime labeled $r_1$). For smaller $\alpha$ (regime labeled $r_{2a}$), the passive region has a certain fixed probability of discovering a beneficial mutation until at least one of the essential bits has mutated. A new round of discovery is started by each copying event from the active to the passive region, and so the discovery rate decreases when $\alpha$, and therefore the number of reset events, is reduced. In between these two extremes, there is a certain optimal rate of gene conversion that maximizes the adaptation rate $r$. It represents the trade-off between resetting the passive region frequently enough via gene conversion from active to passive and, at the same time, allowing enough mutations to occur in between those events to discover a fitter genotype.

A more quantitative understanding of the adaptation by gene conversion in this landscape can be gained by considering the rates of three key processes: First, the rate with which neutral mutations emerge and get fixed in the passive region is $N \mu$, since the mutation rate in the whole population is $N \mu$, and the fixation probability is $1/N$. Second, gene conversions from active to passive are also neutral, and therefore they get fixed with rate $\alpha$. The third key process is the one which actually leads to fitness gains: Gene conversion that transfers a beneficial mutation from the passive to the active region. The rate of this gene conversion is $Na$, but its fixation depends on the state of the passive copy, i.e., the probability $P_b$ that the potential fitness of the passive region is larger than the actual fitness of the population at the time of conversion. If so, the active gene will quickly evolve to a new fitness maximum with a fitness increase $\Delta f$ whose expectation value $\langle \Delta f \rangle$ can be measured numerically from a statistical analysis of our landscapes (see Fig. S4d). If the potential fitness is less than the actual fitness, the passive-active conversion event is discarded.

As long as $Na \gg \mu N$ (which might be the case even if $\alpha < \mu N$ for sufficiently large population size $N$), conversion events are so frequent that potentially beneficial genotypes in the passive region are instantly copied to the active region and get fixed in the population (the passive region is “continuously activated” on the time scale of mutations). Therefore, gene conversion from passive to active is not rate limiting and we can ignore this step.

We can differentiate the following regimes (see File S1 for detailed derivations): (1) For large $\alpha \gg \mu N$, the limiting factor is the ability of the passive region to accumulate $k$ mutations before it is reset again by conversion from the active to the passive region. We know that $k > 1$ is required since the population always resides at a local maximum in the landscape. For small $\mu \ll 1$, the probability of finding a beneficial mutation at $k = 2$ is much greater than for any $k > 2$ so the latter can be neglected, leading to an adaptation rate of

$$r_1 \approx \frac{\langle \Delta f \rangle (\mu N)^2}{\alpha} P_2$$

(we use the suffix $i$ in $r_i$ to identify the approximations of the adaptation rate $r$ in different regimes). Here $P_2$ is the fraction of genotypes at at Hamming distance $k = 2$ from the local maximum which are beneficial (see Fig. S4a). (2) In the opposite limit $\alpha \ll \mu N$, beneficial mutations can be discovered in two ways: (2a) in the wake of a conversion event from active to passive while the essential bits have not yet mutated, and (2b) by randomly discovering a beneficial genotype anywhere in the fitness space long after the conversion event from active to passive, when the initially correct configuration of the essential bits has already been lost. The latter is highly unlikely since it requires that all $N_0$ essential bits are again correct by chance, i.e. are in one out of $2^{\frac{N_0}{\mu}}$ configurations. Therefore, the contribution of (2b) can only become comparable to that of (2a) for very small $\alpha$, when the exploration phase of the passive region is very long to balance this exponentially small probability. Because of the ruggedness of the landscape, we assume that altering any bit (other than the essential bits) has the same fixed probability $P_g$ of yielding a fitter genotype. This probability is equal to the fraction of all genotypes with correct essential bits that have a higher fitness than the current maximum (see Fig. S4a), leading to

$$r_{2a} = \frac{N - N_0}{N_0} P_g \alpha$$

for case (2a). While $r_1$ decreases with $\alpha$ as $1/\alpha$, $r_{2a}$ is a linear function of $\alpha$, thus an optimum speedup indeed occurs at the intermediate conversion rate $\alpha \propto \mu N$. In case (2b), i.e. for random discovery of beneficial genotypes (which are still continuously activated), the rate of fitness increase is

$$r_{2b} = \frac{N \alpha}{2} N_0 2^{-N_0} P_g$$

Note that $r_{2b}$ does not depend on $\alpha$ since the generation of new genotypes in the passive region happens with rate $\mu N$ and any beneficial passive genotype is reliably activated.

The estimates (18),(19), and (20) are shown in Figure 8d superimposed with our numerical results. Figure S5a in addition shows $r$ as a function of the gene conversion rate $\alpha$ for different values of the ruggedness parameter $X$. As can be seen, the approximations (18), (19) and (20) hold remarkably well across all levels of ruggedness.

For even smaller $\alpha \ll \mu N/N$ (regime 3), the passive copy is no longer “continuously activated” (in contrast to all other cases considered above), and therefore many mutations occur in the passive region before the next gene conversion event from the passive to the active region. A very slow fitness gain can also be expected in this regime as a result of random discovery of beneficial genotypes anywhere in the fitness space, with the average speedup rate given by

$$r_3 \approx \frac{N \alpha}{2} N_0 2^{-N_0} P_g$$

(see File S1 and Fig. S5b). We should note however that in reality gene conversion itself might become less likely or cease completely when sequences of the original and the copy diverge sufficiently far (Teshima and Innan 2004). In light of this, the regimes of gene conversion via random discovery (regimes 2b and 3) would likely be suppressed. However, the optimal speedup occurs in the large $\alpha$ regime in which the homology between the original and the copy is well preserved and our assumption of fixed conversion rate is justified.

**Discussion**

The idea that gene duplication might facilitate evolution has been around since the 1970s when Ohno (2013) suggested that duplicate DNA sequences may escape strong selective pressure that usually prevents innovation beyond small incremental steps.
It has also been argued that gene duplication and subsequent deactivation, mutation and reactivation might have been an important mechanism during early evolution (Koch 1972), but consensus emerged that the most probable fate of the new copy is pseudogenization (Lynch and Conery 2000; Innan and Kondrashov 2010). Concerted evolution due to gene conversion only recently entered the picture; however, it has already been suggested as a possible mechanism of accelerated evolution of the NspH gene family in C. elegans (Thomas 2006). A theoretical study by Mano and Innan (2008) explaining this observation investigated the fixation of a single mutation within multiple active copies of a gene subject to gene conversion. Previous studies implicated interlocus gene conversion from non-functional to functional genes as an important molecular mechanism leading to inherited diseases (Chen et al. 2007), but the evolutionary implications of this process have not been thoroughly addressed.

Our study provides the first theoretical description of adaptive evolution of a functional gene on a rugged fitness landscape in the presence of gene conversion with its non-functional duplicate. We demonstrate that gene conversion can significantly speed up adaptation: As long as there is a homologous region in the duplicated sequence that can serve as the template and target of gene conversion events, it can help to explore paths of adaptation in non-monotonous landscapes that would otherwise be dramatically slowed down due to deleterious intermediate steps.

As the comparison across the different regimes in our minimal model shows, gene conversion provides a particular advantage for smaller populations due to stochastic finite size effects that render direct crossing of even shallow fitness valleys unlikely, while deterministically gene conversion only leads to a moderate speed-up for extremely deep fitness valleys. In the example in Figure 7b, the maximum speedup occurs at a population size on the order of 100 individuals. Because of this general trend, one might therefore speculate that the maintenance of pseudogenes for the purpose of optimizing adaptation is more beneficial for higher organisms with relatively small population sizes, as opposed to, e.g., bacteria. It is important to bear in mind though that the advantage of gene conversion is lost when the selective disadvantage of the intermediate mutants gets very small (equivalently, the population size may become so small that the fitness valley is effectively flat). In this case, adaptation cannot benefit from the existence of the gene conversion path as the additional mutation required slows its exploration.

In our modeling, we did not consider the process of gene duplication that originally led to the creation of the copy, neither did we consider the possibility of elimination of the duplicate gene. Consequently, we did not specify how the passive copy is freed from selective pressure, but since the “gene” we consider might actually be only a part of a larger, functional gene, there are a variety of possible mechanisms which would render the copy inactive and still leave a large enough region intact for gene conversion. These mechanisms include pseudogenization (of the larger gene) via premature stop codons, heterochromatin silencing, or mutations in cis regulatory regions of the copy. Our results are general in that they do not depend on this mechanism.

Across all dynamical regimes in the minimal model as well as in the rugged landscape model, we find an optimal rate of gene conversion maximizing the adaptation rate of the original gene (note, however, that this optimum does not always lead to an improvement over ordinary adaptation of the original gene, cf. Figure 7b). In both cases, this optimum is due to a trade-off: On the one hand, the passive copy must be able to explore the genetic space, which is only possible for small gene conversion rates a, when the passive copy is not reset to wild-type too frequently. On the other hand, a beneficial mutation must be present and then transferred to the active region, which requires a sufficiently large a.

The first requirement is qualitatively similar in both our models: For too large gene conversion rates a, the passive gene gets stuck in the state of the active gene. However, the second requirement (the reason why too small gene conversion rates are also sub-optimal) is due to different reasons in our two models: While in the NK-landscape model, the accumulation of deleterious mutations prohibits finding beneficial genotypes at small gene conversion rates (as only the reset to the active version is likely to reverse all deleterious mutations in the passive region), the minimal model only considers two loci and ignores other mutations. In the latter, a large enough gene conversion rate is required simply because the beneficial mutation needs to be transferred to the active region. This is usually not a concern in the NK-landscape model: if the gene conversion rate a is on the order of the total mutation rate of any bit N (the region of the optimum in between approximations (18) and (19)), this automatically implies Na ≫ μN for sufficiently large populations leading to reliable transfer to the active region. Even though the fixation of beneficial mutations in the passive region might be rare without the constraint of natural selection, their reactivation in successful cases would be very likely. Thus, together, our two models point to two potential reasons why a sufficiently large the gene conversion rate is required for efficient adaptation.

Note also that we assumed equal conversion rates a in both directions, whereas in reality the underlying molecular mechanisms and the genetic context could potentially lead to differences between the forward and the backward rates. In addition, gene duplication itself may also contribute to the forward rate of resetting the passive region to the original genotype, as it produces a fresh copy that can get silenced and become a new template for gene conversion as described by our model.

The quantitative applicability of our findings to experimental evolution is limited by our choice of model fitness landscapes. Natural fitness landscapes may have quite different structure than abstract NK-type landscapes, which could affect our estimates. For example, if finding a beneficial genotype required k > 2 single mutations, the adaptation rate r1 (in the regime where resetting of the silent copy is rate limiting) would be proportional to (μN)k; however, the optimal conversion rate would still scale as μN. Our theory also assumed that the probability of yielding a fitter genotype, which we used to obtain Equation (19), is the same for any mutation and any local fitness peak. In reality, the fitness landscape parameters may vary greatly. Nonetheless, we believe that the salient qualitative features of evolutionary dynamics accelerated by gene conversion, including the presence of an optimal conversion rate commensurate with the mutation rate, are robust and can be expected to play an important role in adaptive evolution.

Data availability

The authors state that all information necessary for confirming the conclusions presented in the article is represented fully within the article and the Supplemental Material.
Acknowledgements

We are grateful to Michael McLaren for his invaluable comments on the manuscript, and to Anne Carvunis and Sergey Kryazhimskiy for stimulating discussions. This work was supported by the NSF grant MCB-1616997 and the San Diego Center for Systems Biology, NIH/NIGMS grant P50-GM085764. P.B. acknowledges support from HFSP fellowship LT000840/2014-C.

Literature Cited


