Spiraling Complexity: A Test of the Snowball Effect in a Computational Model of RNA Folding

Ata Kalirad* **and Ricardo B. R. Azevedo***

*Department of Biology and Biochemistry, University of Houston,, Houston, Texas, United States of America

File S1: Derivation of Equations 7 and 8

An inviable introgression from one lineage to another may be caused by one or more DMIs. If we assume that all inviable single introgressions are caused by simple DMIs, we can apply the Orr model.

Without loss of generality, we assume that all substitutions take place in one lineage and that introgressions are assayed in either the derived → ancestral or the ancestral → derived direction. Below, we take each case in turn.

Derived → ancestral

Introgressions of derived alleles can be treated independently after each substitution. The change in the number of derived inviable single introgressions following substitution $k+1$ is given by

$$\Delta J = J_{k+1} - J_k = \begin{cases} 0\,, & \text{if} \quad \Delta I = 0 \\ 1\,, & \text{if} \quad \Delta I \geqslant 1 \end{cases}$$

For simplicity we refer to $J_k^{(1)}$ as J_k.

Following the $k+1$ substitution, the Orr model assumes that ΔI simple DMIs arise from k independent Bernoulli trials with probability of success p. Therefore, $\Delta J = 0$ is expected to occur with probability $(1-p)^k$, and $\Delta J = 1$ with probability $1-(1-p)^k$.

Thus, the expected number of derived inviable introgressions is given by

$$J_{k+1} = J_k + 1 - (1-p)^k \tag{S1}$$

Assuming $J_1 = 0$, the solution to difference Equation S1 is

$$J_k = k - \frac{1-(1-p)^k}{p} \quad . \tag{S2}$$

Equations S1 and S2 are Equations 7 and 8 in the main text, respectively.

Ancestral → derived

Calculating the number of inviable ancestral introgressions is complicated by the fact that each ancestral allele can be "recruited" into a DMI after each substitution occurring after its corresponding derived allele has substituted.

After the k substitution, there are J_k ancestral inviable introgressions, and $k - J_k$ ancestral alleles that, when introgressed, do not cause inviability. Thus, the expected number of ancestral inviable introgressions is given by

$$J_{k+1} = J_k + kp\left(1 - \frac{J_k}{k}\right) = J_k(1-p) + kp \tag{S3}$$

Assuming $J_1 = 0$, the solution to difference Equation S3 is also Equation S2.

Conclusion

Equations S1 and S2 describe the accumulation of all inviable single introgressions in the Orr model, regardless of whether the introgressed alleles are derived or ancestral.

File S2: Detecting Simple DMIs

Here we use the general terms genotypes, loci and alleles, instead of sequences, sites and nucleotides.

Two viable genotypes, 1 and 2, differ at $k \geqslant 2$ loci. Loci are denoted by $A, B, C, \ldots$ The alleles of genotype 1 are indicated by a subscript 1 $(A_1, B_1, C_1, \ldots)$; the alleles of genotype 2 are indicated by a subscript 2 $(A_2, B_2, C_2, \ldots)$. Introgression of the A_1 and B_1 alleles from genotype 1 to genotype 2 is denoted $1 \xrightarrow{A,B} 2$.

There is a simple DMI between the A_1 and B_2 alleles if *all* of the following 6 conditions are met.

1. The single introgression $1 \xrightarrow{A} 2$ results in an inviable genotype. On its own, this condition indicates that there is a DMI between the A_1 allele and one or more alleles from genotype 2 at the remaining $k - 1$ loci $(B_2, C_2, \ldots)$.

2. The single introgression $2 \xrightarrow{B} 1$ results in an inviable genotype. On its own, this condition indicates that there is a DMI between the B_2 allele and one or more alleles from genotype 1 at the remaining $k - 1$ loci $(A_1, C_1, \ldots)$. Taken together, conditions #1–2 are not sufficient to indicate that the A_1 and B_2 alleles participate in the same DMI.

3. The double introgressions $1 \xrightarrow{A,B} 2$ and $2 \xrightarrow{A,B} 1$ both result in viable genotypes. In other words, a second introgression rescues viability. Taken together, conditions #1–3 indicate that the A_1 and B_2 alleles participate in the same DMI; the conditions do not, however, rule out the possibility that the DMI involves additional alleles from either genotype at the remaining $k - 2$ loci $(C, D, \ldots)$. In other words, the DMI might be simple or complex.

4. A_1 and B_2 are not both ancestral. If conditions #1–3 are met but condition #4 is violated, then the DMI must involve a derived allele at an additional locus—i.e., the DMI is complex—because A_1 and B_2 were not incompatible in the ancestor.

5. If both A_1 and B_2 are derived alleles, this condition is ignored. If A_1 is an ancestral allele, then the B_2 substitution occurred after the A_2 substitution; if B_2 is an ancestral allele, then the A_1 substitution occurred after the B_1 substitution. If conditions #1–4 are met but condition #5 is violated then the DMI is complex because A_1 and B_2 were not incompatible in the background in which the derived allele arose.

6. If the latest substitution at either the A or the B locus was the i-th substitution, and $i < k$, then conditions #1–3 are also met in all genotypes present after the i-th substitution. If conditions #1–5 are met but condition #6 is violated then the DMI is complex because A_1 and B_2 were not incompatible in some genetic backgrounds.

To count simple DMIs in our simulations, we introgress nucleotides between the two sequences at each of the k divergent sites, in both directions. Every time an introgression results in an inviable genotype (condition #1), we look for another introgression in the opposite direction that also results in an inviable genotype (condition #2). We then test both double introgressions involving these alleles to test for condition #3. If we find a pair of alleles satisfying conditions #1–3, we test for conditions #4–6 directly.

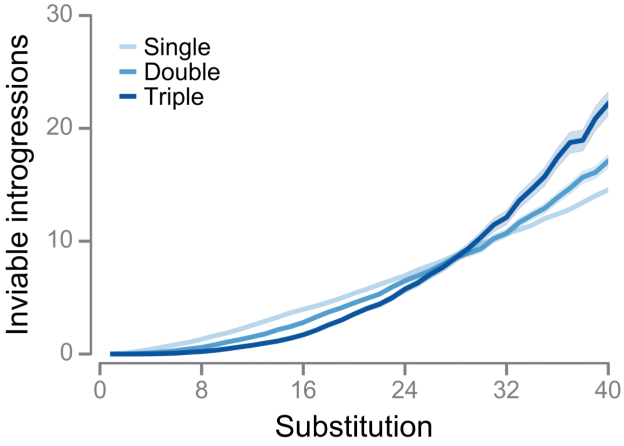

Figure S1 Allowing sites to undergo multiple substitutions does not affect the pattern of accumulation of inviable single, double, and triple introgressions (Figure 2). Values are means of 10^3 simulations with $\alpha = 12$. Shaded regions indicate 95% confidence intervals, CIs.

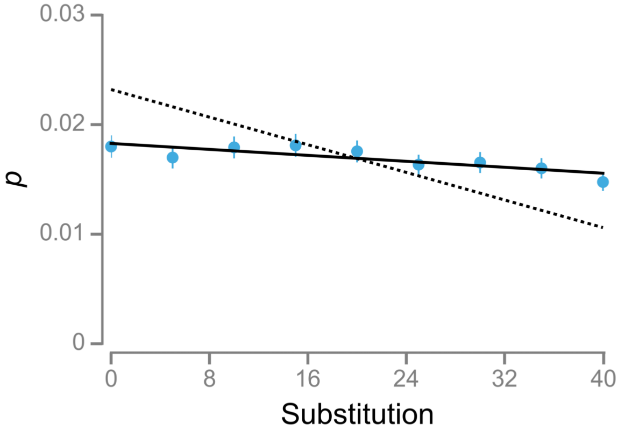

Figure S2 The probability that a simple DMI appears is approximately constant in the RNA model. We measured p directly in each simulation in the ancestor and in one lineage at $k = 5, 10, 15, \ldots, 40$ (Equation 5). Values are means and 95% CIs of 10^3 simulations with $\alpha = 12$. The solid line is a linear regression fit on the mean values of p. The dashed line is the pattern of decline of p that would be expected to generate a trend in the number of inviable single introgressions most similar to that in Figure 2 using Equation 7.

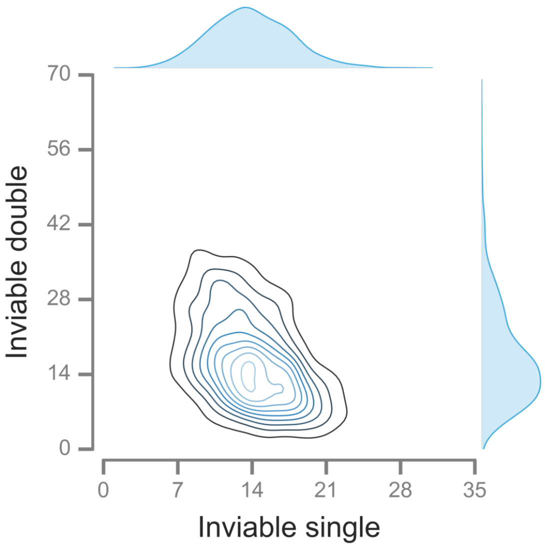

Figure S3 The numbers of inviable single and double introgressions after $k = 40$ substitutions are negatively correlated with each other in the RNA model ($\rho = -0.469, P < 10^{-6}$). One- and two-dimensional kernel density estimates based on 10^3 simulations with $\alpha = 12$.

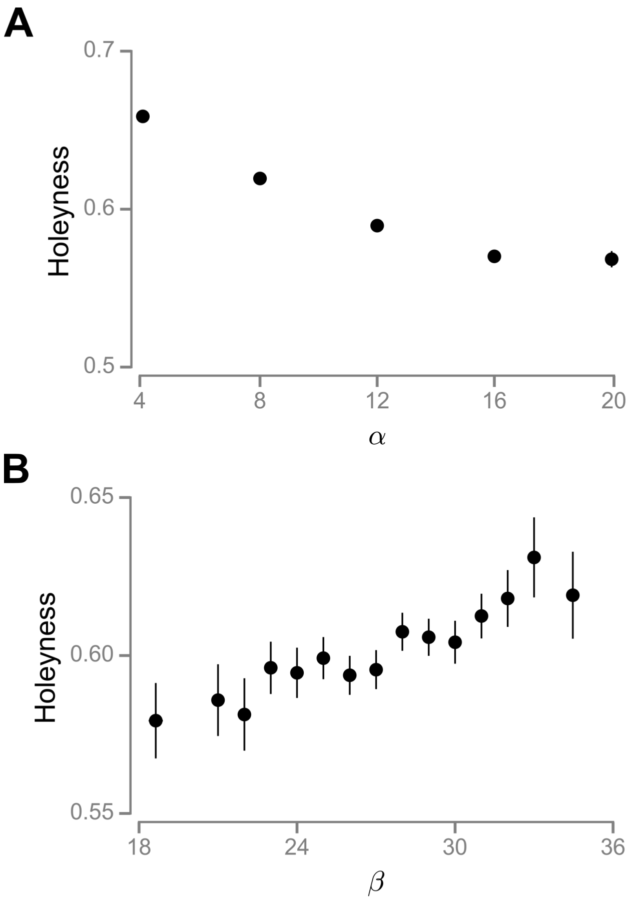

Figure S4 Holeyness decreases with the value of α (A) and increases with the number of base pairs, β, in the reference sequence (B). (A) Values are means of 10^3 RNA folding simulations for each value of α. (B) The holeyness data from the 5×10^3 simulations used in (A) were grouped by individual values of β. We pooled estimates for $\beta \leqslant 20$ and for $\beta \geqslant 34$. The resulting β groups have sample sizes ranging from 125 to 552. Error bars are 95% CIs. The error bars in (A) are covered by the points.

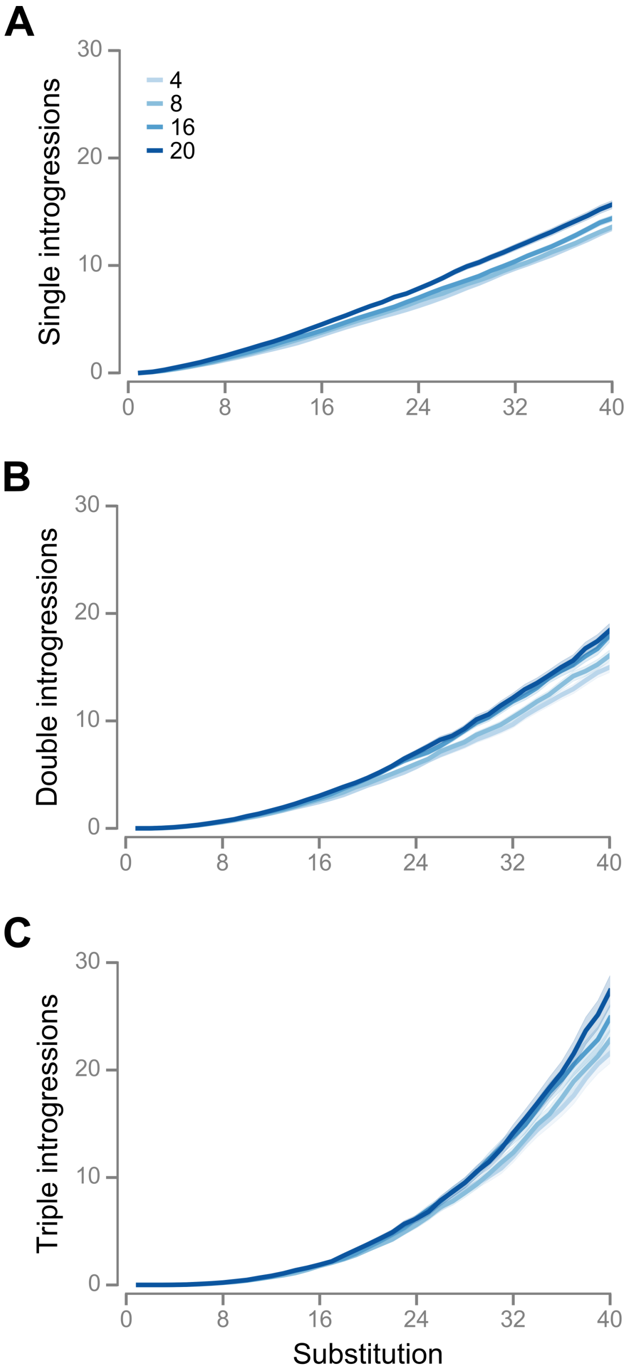

Figure S5 Effect of α on the accumulation of inviable single (A), double (B), and triple (C) introgressions in the RNA model. Values are means of 10^3 simulations for each value of α. Shaded regions indicate 95% CIs.

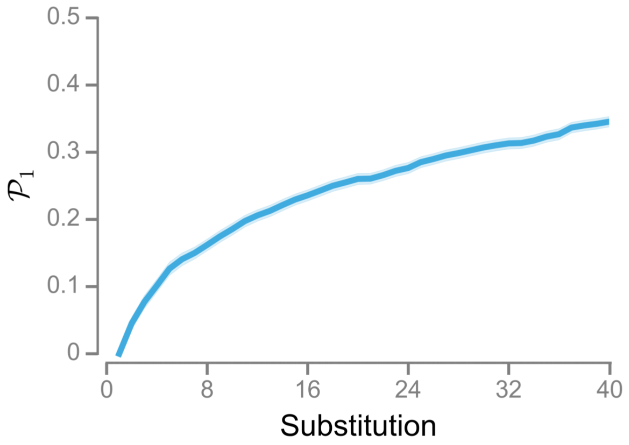

Figure S6 Evolution of the proportion of single inviable introgressions, $\mathcal{P}_1$, as populations diverge in the RNA model. Values are means of 10^3 simulations with $\alpha = 12$. Shaded region indicates 95% CIs.

Table S1 Properties of the 10^3 ancestors used in the simulations with $\alpha = 12$

Property	Mean	(Standard deviation)
Sequence		
GC content	0.52	(0.05)
Hamming distance from reference sequence	55.94	(5.09)
Structure		
Minimum free energy (kcal mol^{-1})	–24.98	(6.01)
Number of base pairs	25.63	(3.99)
Holeyness	0.58	(0.12)
Base pair distance from the reference sequence	11.24	(1.09)
Ensemble		
Base pair distance between pairs of sequences	50.93	(5.79)

Table S2 Estimates of the parameters in Equation 6. The model was fitted by nonlinear least-squares regression to the average numbers of inviable single, double, and triple introgressions shown in Figure S1.

Introgressed alleles, i	DMI order, n	a_i	(95% CI)	b_i	(95% CI)	R^2
1	$\geqslant 2$	0.094	(0.003)	1.37	(0.0098)	0.999
2	$\geqslant 3$	0.024	(0.002)	1.80	(0.026)	0.999
3	$\geqslant 4$	0.0036	(0.0004)	2.42	(0.036)	0.998

Table S3 Estimates of the parameters in Equation 6. The model was fitted by nonlinear least-squares regression to the average numbers of inviable single, double, and triple introgressions shown in Figure S5.

Introgressed alleles, i	DMI order, n	a_i	(95% CI)	b_i	(95% CI)	R^2
$\alpha = 4$						
1	$\geqslant 2$	0.075	(0.002)	1.41	(0.010)	0.999
2	$\geqslant 3$	0.023	(0.001)	1.78	(0.016)	0.999
3	$\geqslant 4$	0.004	(0.0004)	2.35	(0.026)	0.999
$\alpha = 8$						
1	$\geqslant 2$	0.097	(0.003)	1.35	(0.008)	0.999
2	$\geqslant 3$	0.022	(0.001)	1.81	(0.017)	0.999
3	$\geqslant 4$	0.003	(0.0002)	2.49	(0.022)	0.999
$\alpha = 16$						
1	$\geqslant 2$	0.10	(0.003)	1.35	(0.009)	0.999
2	$\geqslant 3$	0.025	(0.002)	1.80	(0.019)	0.999
3	$\geqslant 4$	0.003	(0.0004)	2.45	(0.031)	0.999
$\alpha = 20$						
1	$\geqslant 2$	0.13	(0.004)	1.31	(0.009)	0.999
2	$\geqslant 3$	0.027	(0.001)	1.79	(0.015)	0.999
3	$\geqslant 4$	0.002	(0.0002)	2.61	(0.035)	0.999