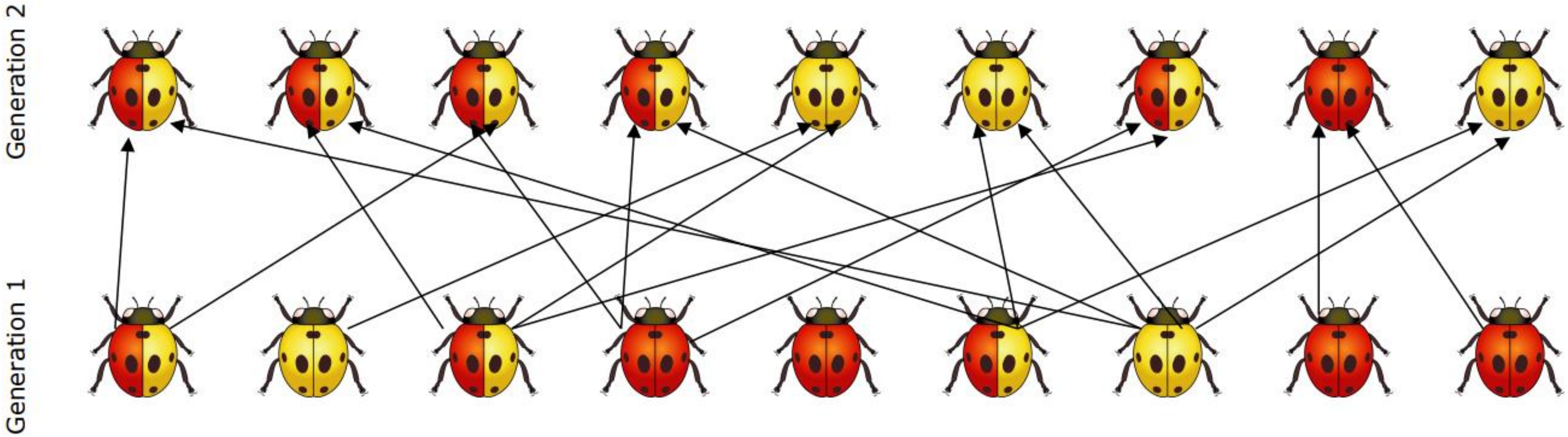
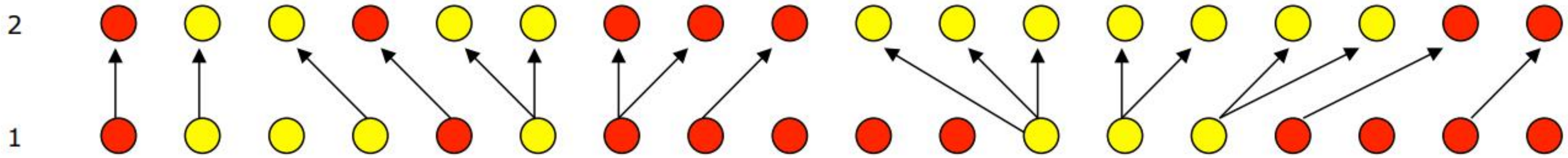
遗传漂变

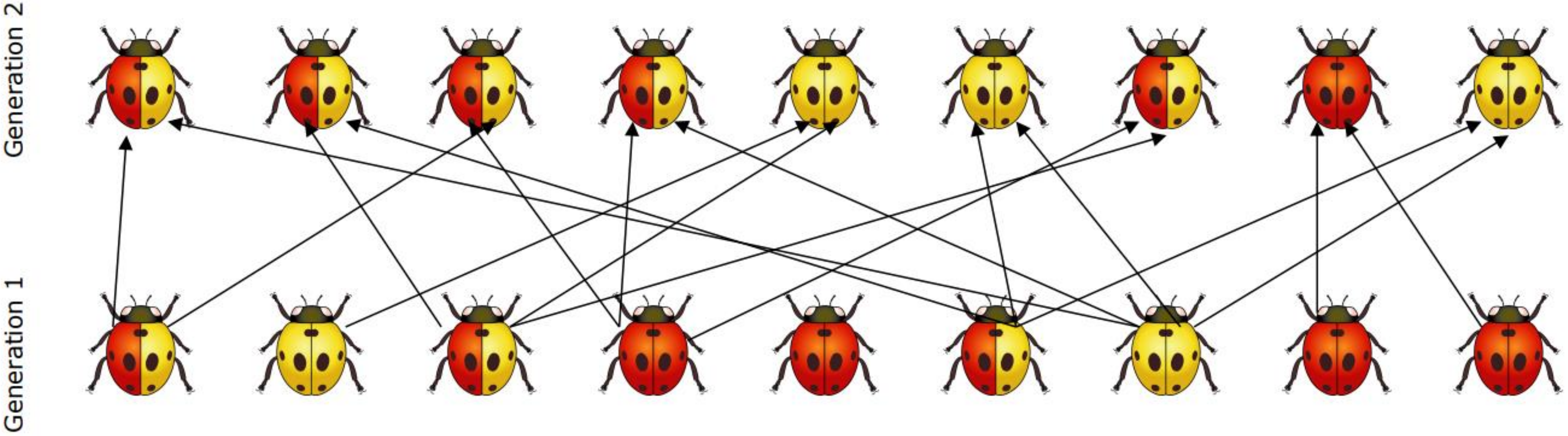
周华瑞

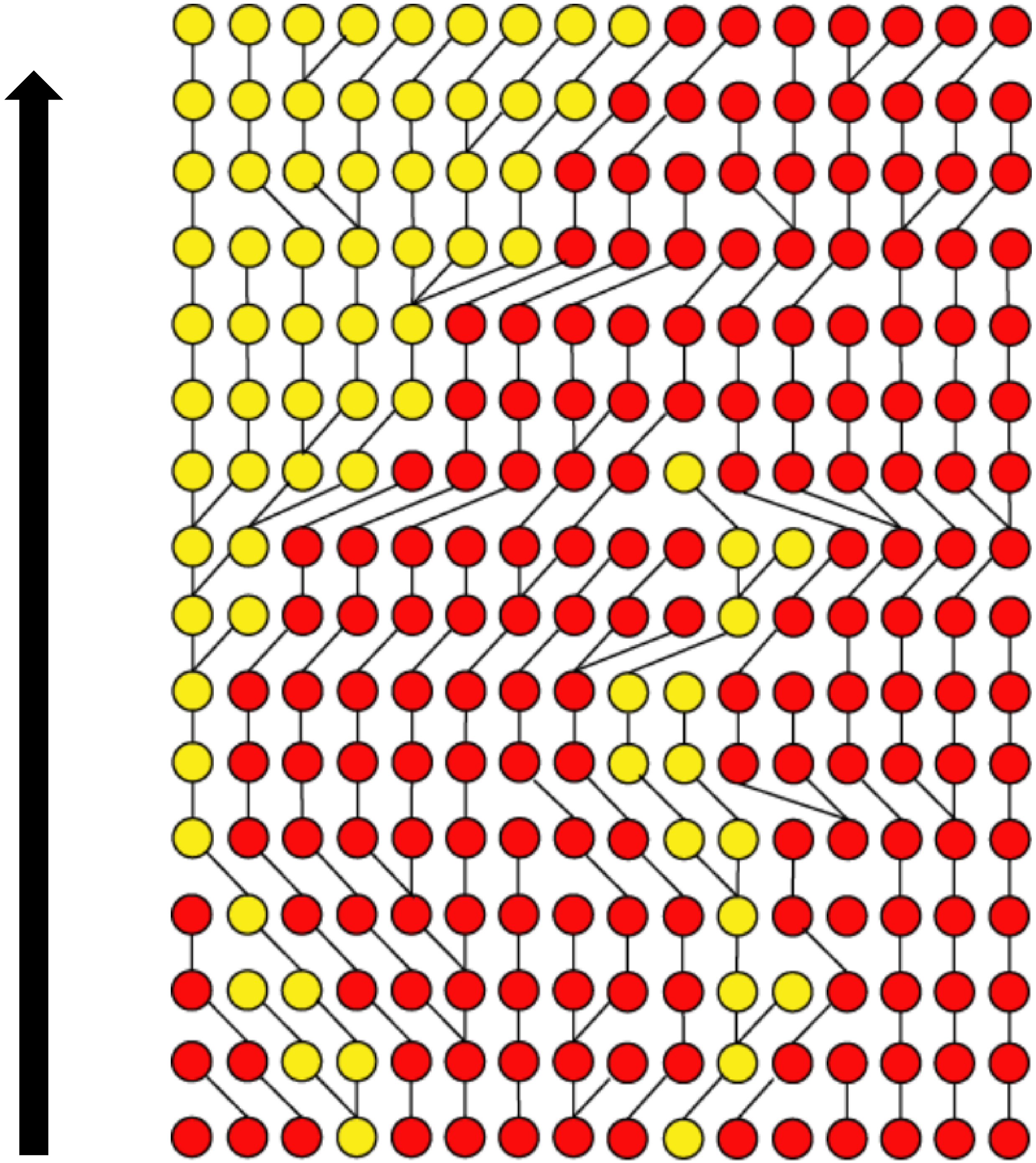
Wright-Fisher Model

- 假设有一个大小为N的种群，记为G_0，每个个体有两个A基因位点，A基因具有两个等位基因A1和A2. G_0的下一代G_1的种群大小仍为N，并且G_1中的每一个基因位点都随机来自于G_0，即对于每一个G_1个体，随机从2N个G_0代的A基因中抽取两个(放回抽样)构成自己的基因. 不考虑突变和自然选择.

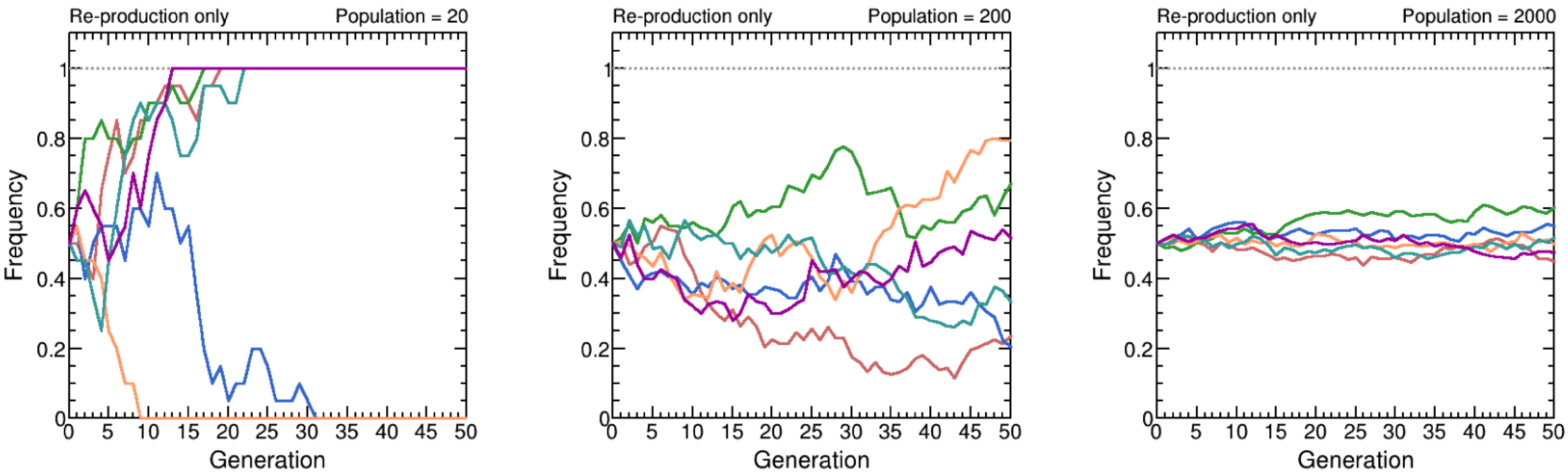



2
1


Generation 2
Generation 1

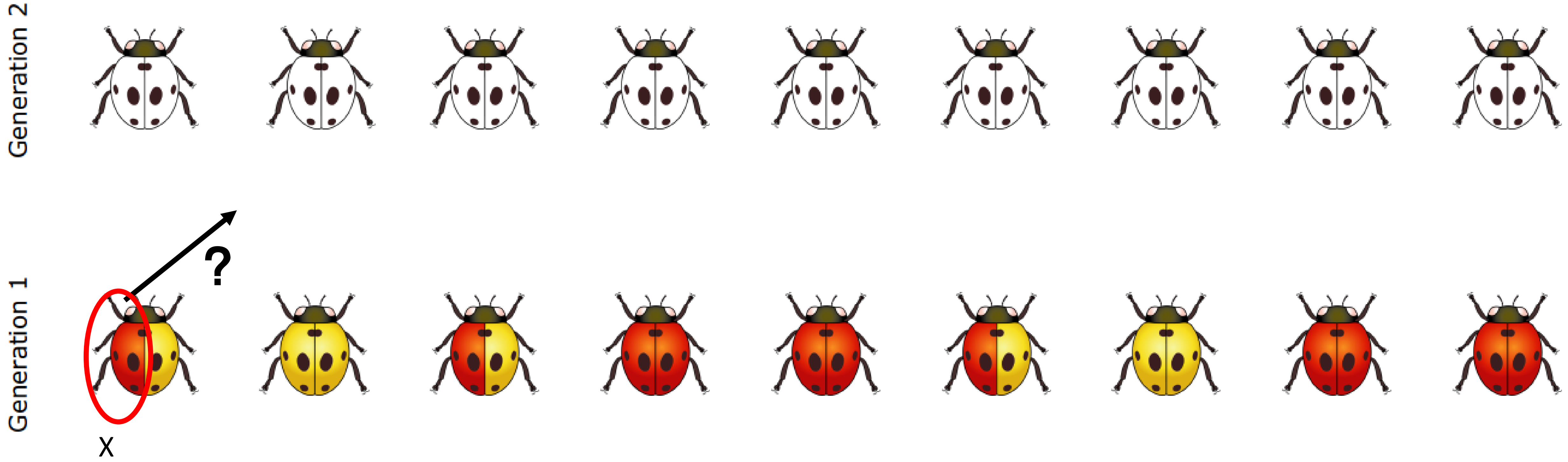


小种群更容易出现基因的丢失



https://boundino.github.io/S188592web/drift.html

问题1: G_0代某一个基因位点进入G_1的概率是多少？



问题1: G_0代某一个基因位点进入G_1代的概率是多少?

- 考虑G_1代的一个位点，这个位点来自于G_0代位点X的概率为1/2N, 不来自于位点X的概率为1-1/2N. 由于G_1代共有2N个位点，并且彼此之间独立，因此G_1代所有位点都不来自于X的概率为

$$\left(1-\frac{1}{2N}\right)^{2N}$$

从而位点X进入G_1代的概率为

$$1-\left(1-\frac{1}{2N}\right)^{2N} \to 1-\frac{1}{e} \approx 0.63, \qquad \text{as N} \to \infty$$

例子：某种群有10,000个基因位点，问10代之后还有多少个位点存在？

Answer: $10{,}000 \times 0.63^{10} \approx 98$

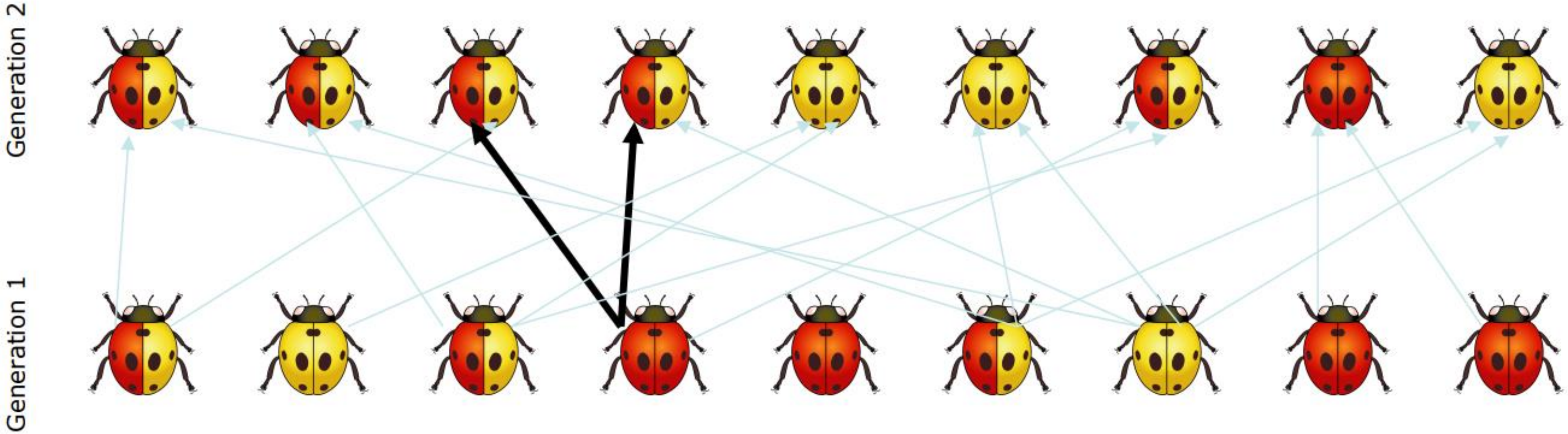
问题2: 已知G_0代A1基因的数目为k，那么G_1代A1基因数目的分布是什么?期望是多少?

$$\mathrm{p}_{km}= P(X_1 = m|X_0 = k) = C_{2N}^{m}\left(\frac{k}{2N}\right)^{m}\left(1-\frac{k}{2N}\right)^{2N-m}$$

$$m = 0, 1, \cdots, 2N$$

$$E(X_1|X_0 = k) = k$$

问题3: 求G_1中随机的两个位点来自于同一个G_0中的位点的概率



第一个位点来自上一代祖先的概率为1，第二个位点来自同一个祖先的概率为1/2N，因此两个位点来自同一个G_0代祖先的概率为 $\frac{1}{2N}$

问题3-1: 求G_t中任意的两个位点来自于G_0中同一个位点的概率

- 先求G_t中任意的两个位点不来自于G_0中同一个位点的概率，即

$$\left(\mathbf{1}-\frac{\mathbf{1}}{\mathbf{2}\mathrm{N}}\right)^{\mathrm{t}}$$

从而来自于G_0中同一个位点的概率为

$$\mathbf{1}-\left(\mathbf{1}-\frac{\mathbf{1}}{\mathbf{2}\mathrm{N}}\right)^{\mathrm{t}}$$

问题3-2: 求G_t中任意的两个位点来自于G_0中的同一个位点，且不来自于G_1中同一个位点的概率

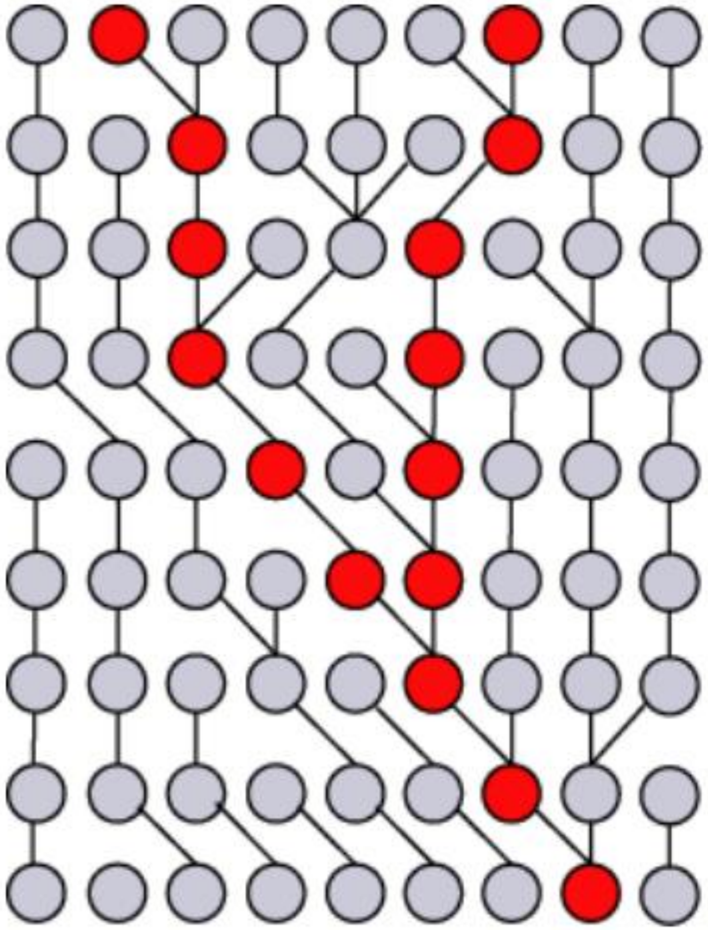
首先这两个位点不来自于G1中同一个位点的概率为

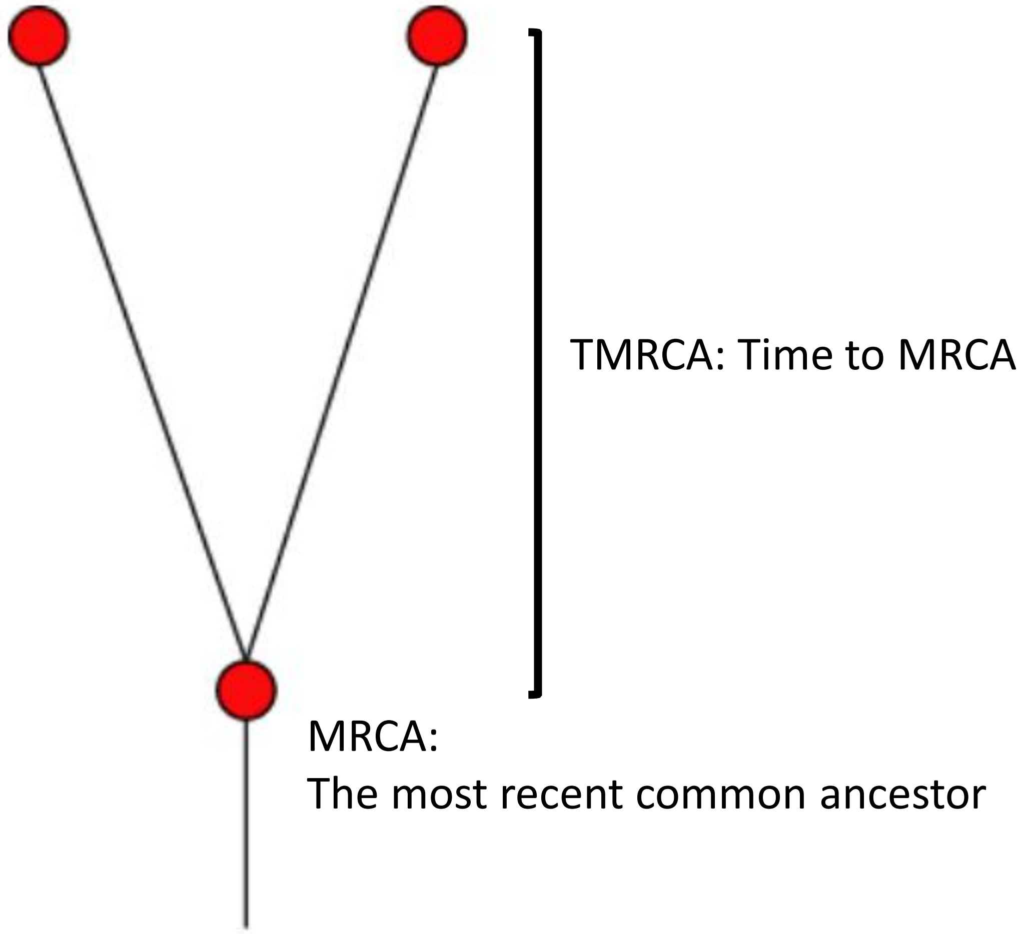
$$\left(\mathbf{1}-\frac{\mathbf{1}}{\mathbf{2N}}\right)^{\boldsymbol{t-1}}$$

然后G1中两个位点来自于同一个G0中位点的概率为1/2N,
从而题目所求概率为

$$\frac{\mathbf{1}}{\mathbf{2N}}\left(\mathbf{1}-\frac{\mathbf{1}}{\mathbf{2N}}\right)^{\boldsymbol{t-1}}$$

合并理论(Coalescence theory)



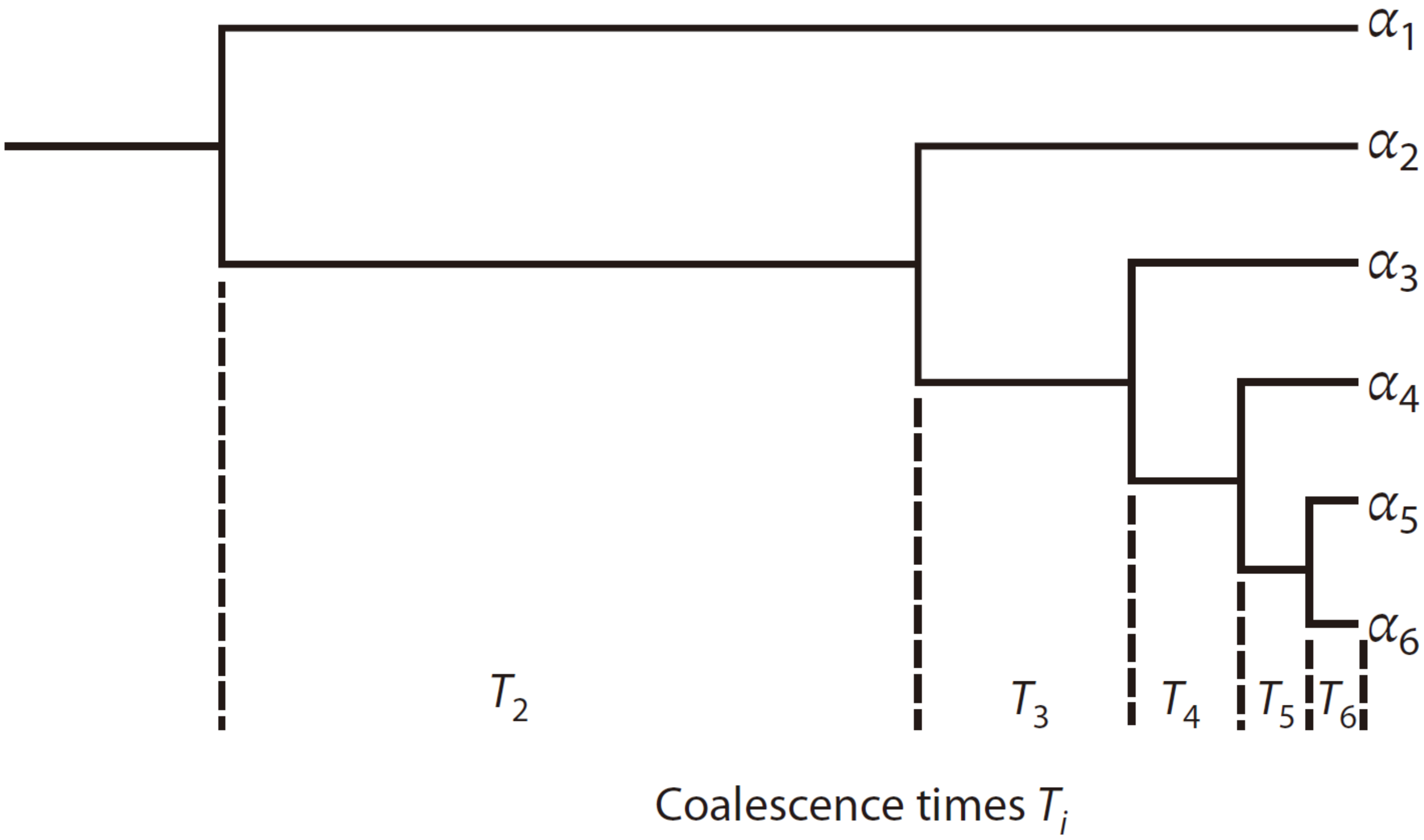


G_t中两个位点的平均TMRCA（假设种群大小不变）

- $P(T_{MRCA} = 1) = \frac{1}{2N}$
- $P(T_{MRCA} = 2) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)$
- $P(T_{MRCA} = k) = \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^{k-1}$
- …
- $E(T_{MRCA}) = \sum_k^\infty kP(T_{MRCA} = k) = 2N$

这实际上是一个几何分布

G_t中n个位点的平均TMRCA （假设种群大小不变）



A primer of population genetics and genomes, 4th edition.

G_t中n个位点的平均TMRCA （假设种群大小不变）

我们先来求n个位点的平均合并时间(Coalescence time， T_C)，即发生合并事件的时间, 注意n个位点在前一代不发生合并的概率为

$$\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{n-1}{2N}\right)$$

因此在前一代发生至少一次合并事件的概率为

$$1-\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{n-1}{2N}\right)$$

$$\approx\frac{1+2+\cdots+(n-1)}{2N}=\frac{n(n-1)}{4N}$$

- $P(T_{C;n} = 1) = \frac{n(n-1)}{4N}$
- $P(T_{C;n} = 2) = \frac{n(n-1)}{4N}\left(1 - \frac{n(n-1)}{4N}\right)$
- $P(T_{C;n} = k) = \frac{n(n-1)}{4N}\left(1 - \frac{n(n-1)}{4N}\right)^{k-1}$
- …
- $E(T_{C;n}) = \sum_{k}^{\infty} kP\left(T_{C;n} = k\right) = \frac{4N}{n(n-1)}$

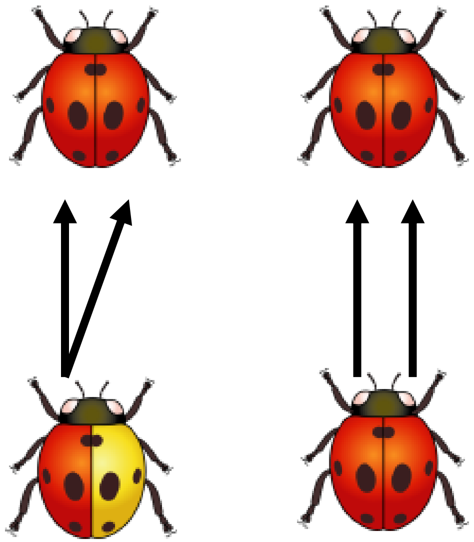
- 从而$E(T_{MRCA}) = \sum_{i=2}^{n} E\left(T_{C;i}\right) = \sum_{i=2}^{n} \frac{4N}{i(i-1)} = \frac{4N(n-1)}{n}$

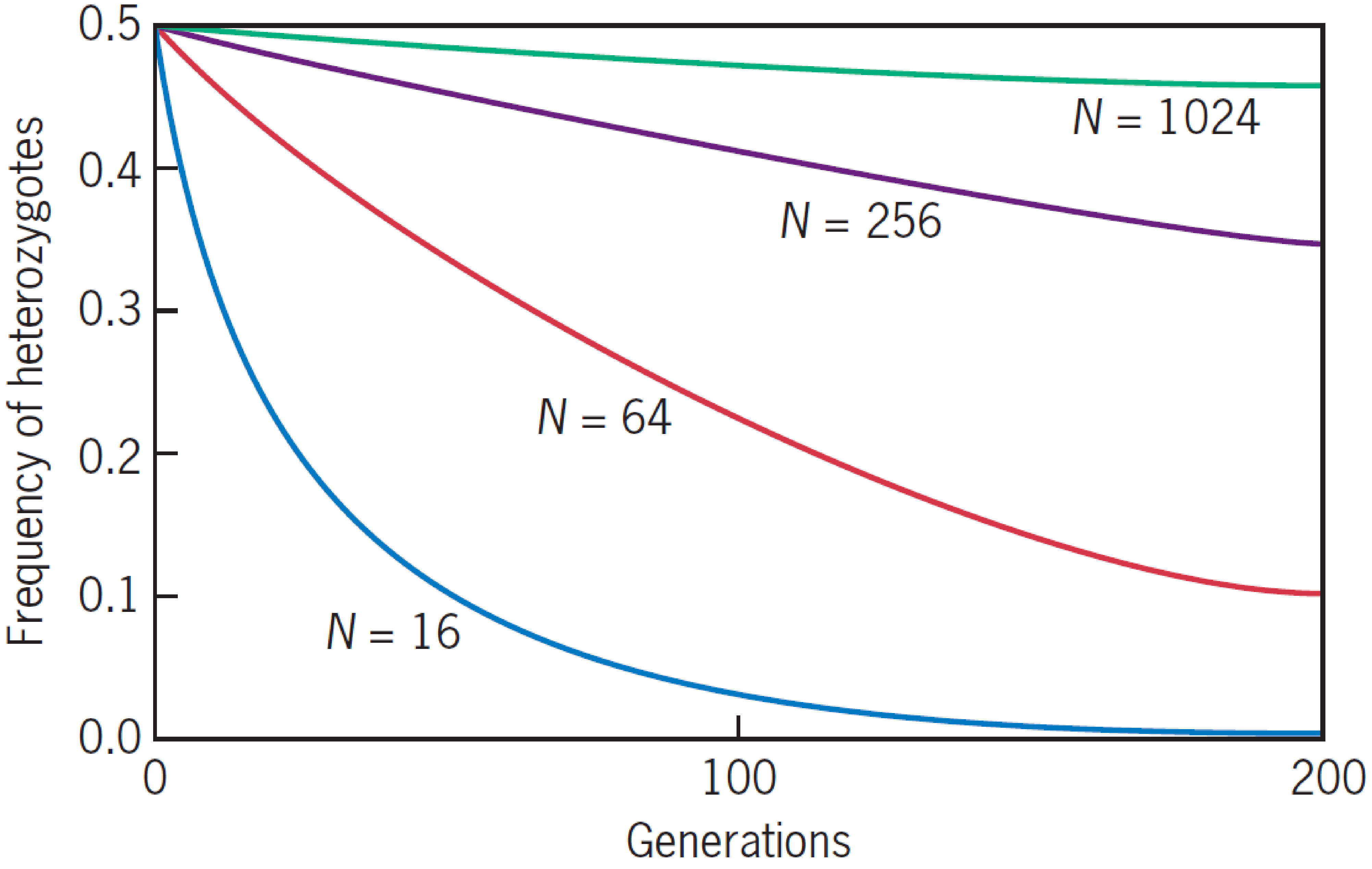
问题4：杂合子的损失

- G_0代有N个个体，假设其中杂合子占比为$\boldsymbol{H_0}$，问G_t代随机选取一个个体是杂合子的概率$\boldsymbol{H_t}$.
- 设G_0纯合子的占比为$\boldsymbol{K_0 = 1 - H_0}$，则我们有

$$\boldsymbol{K_1 = \frac{1}{2\mathrm{N}} + \left(1 - \frac{1}{2N}\right) K_0}$$

从而 $\boldsymbol{H_1 = 1 - K_1 = \left(1 - \frac{1}{2N}\right) H_0, H_t = \left(1 - \frac{1}{2N}\right)^t H_0}$





Frequency of heterozygotes
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
Generations
N = 1024
N = 256
N = 64
N = 16

引入突变

- 设A1以概率u突变成为A1, 求每代杂合子的占比

$$\boldsymbol{K_1 = (1-u)^2\left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)K_0\right)}$$

$$\boldsymbol{\approx (1-2u)\left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)K_0\right)}$$

$$H_1 = 1 - K_1 = 1 - (1-2u)\left(\boldsymbol{\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)(1 - H_0)}\right)$$

$$\boldsymbol{= \left(1 - \frac{1}{2N}\right)H_0 + 2u(1 - H_0)}$$

$$\mathbf{\Delta H = H_1 - H_0 = -\frac{1}{2N}H_0 + 2u(1 - H_0)}$$

令ΔH=0，得到

$$\boldsymbol{H_0 = \frac{4Nu}{1 + 4Nu}}$$

固定和丢失

- 如果在某一代某一种等位基因完全替代了另一个等位基因，例如只有A1没有A2，我们称A1发生了固定(fixation), 而A2发生了丢失(loss).
- 固定和丢失都是吸收态(absorption state)，即一旦发生，后代将保持这种状态.

问题5：设G_0代等位基因A1有i个，求最终A1发生固定的概率

- 由于种群数量有限，而代数无限，因此在若干代之后，G_0代的2N个基因位点只会有一个保存下来，即发生了固定，
- 每一个位点发生固定的概率为1/2N
- 位点和位点之间相互独立，现在有i个A1位点，因此A1最终被固定的概率为i/2N